2018

Recurrence Quantification Models of Human Conversational Grounding Processes: Informing Natural Language Human-Computer Interaction

Clayton D. Rothwell
Wright State University

RECURRENCE QUANTIFICATION MODELS OF HUMAN CONVERSATIONAL GROUNDING PROCESSES: INFORMING NATURAL LANGUAGE HUMAN-COMPUTER INTERACTION

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

By

CLAYTON D. ROTHWELL
B.A., Belmont University, 2007
M.A., Southern Baptist Theological Seminary, 2009
M.S., Wright State University, 2014

2018
Wright State University

Wright State University
GRADUATE SCHOOL

April 13, 2018

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Clayton D. Rothwell ENTITLED Recurrence Quantification Models of Human Conversational Grounding Processes: Informing Natural Language Human-Computer Interaction BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Valerie L. Shalin, Ph.D.
Dissertation Director

Scott N. J. Watamaniuk, Ph.D.
Graduate Program Director

Debra Steele-Johnson, Ph.D.
Chair, Department of Psychology

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

Final Examination

Scott N. J. Watamaniuk, Ph.D.

Ion Juvina, Ph.D.

William S. Horton, Ph.D.

ABSTRACT

Rothwell, Clayton D. Ph.D., Department of Psychology, Wright State University, 2018. *Recurrence Quantification Models of Human Conversational Grounding Processes: Informing Natural Language Human-Computer Interaction.*

Human-human communication is a coordinated dance (Clark, 1996) that requires each participant to consider the other participants. The majority of this coordination centers on the *conversational grounding* process that develops and maintains the common ground, or shared understanding between the individuals (Clark and Schaefer, 1989). Conversational grounding is also a crucial process for human-computer interaction using language-based methods, such as spoken dialogue systems. Previous work has tied grounding processes to the performance outcomes in collaborative tasks (Reitter and Moore, 2014; Gergle et al., 2013, 2004; Clark and Krych, 2004), making it a high priority for increasing capabilities of spoken dialogue systems.

The model of grounding for human-computer interaction should be informed by human-human dialogue. However, the processes involved in human-human grounding are under dispute within the research community. Three models have been proposed: *alignment*, a simple model that has been influential on dialogue system development, *interpersonal synergy*, an automatic coordination emerging from interaction, and *audience design*, a strategic interaction based on intentional coordination. Interpersonal synergy and audience design are two different types of coordination models.

Previously, only one study has tested both the alignment and coordination models simultaneously. Fusaroli and Tylén (2016) introduced communication models based on recurrence quantification analysis to model the amount of repetition between speakers. The current research extended their models to differentiate between the types of coordination. Throughout, the current research applied Fusaroli and Tylén's methods to richer stimuli/-tasks that generate longer dialogues with larger vocabulary and more influences on performance outcomes. Through analysis of four different dialogue tasks, the current work also

examined how common ground processes change as a function of the task characteristics. Subsequent analyses investigated the validity of the nascent recurrence models.

The results showed strong support for the coordination model over the alignment model in human-human communication, Additional results suggested that coordination is the audience design variant. These results place a new requirement on the design of human-computer interaction mechanisms.

Contents

1 Introduction **1**
- 1.1 Human Grounding Process . . . 5
- 1.2 Process-Outcome Relationship and Task Characteristics . . . 6
- 1.3 Common Ground Mechanisms . . . 9
  - 1.3.1 Alignment . . . 9
  - 1.3.2 Coordination . . . 10
  - 1.3.3 Previous Research Investigating Common Ground Models . . . 13
- 1.4 Current Research . . . 16
- 1.5 Dissertation Organization . . . 17
- 1.6 Quantification of Recurrence . . . 17
  - 1.6.1 Recurrence Analysis Illustration . . . 18
  - 1.6.2 Grounding Process Models . . . 21
- 1.7 Extending Recurrence Models . . . 23
  - 1.7.1 Examining Validity . . . 25
- 1.8 Informing Dialogue Systems . . . 25
  - 1.8.1 Task Management and Articulation . . . 26
  - 1.8.2 Clarification Dialogues in Dialogue Systems . . . 28
  - 1.8.3 Measures of Grounding . . . 30
- 1.9 Summary of Hypotheses . . . 31

2 Method **32**
- 2.1 Materials . . . 32
  - 2.1.1 Map Task Corpus . . . 33
  - 2.1.2 Uncertainty Elicitation Task Corpus . . . 35
  - 2.1.3 Diapix Task Corpus . . . 40
  - 2.1.4 Combat Search and Rescue (CSAR) Task Corpus . . . 43
- 2.2 Recurrence Analyses . . . 48
  - 2.2.1 Categorical Data . . . 48
  - 2.2.2 Continuous Data . . . 50
  - 2.2.3 Recurrence Parameters . . . 51
  - 2.2.4 Track 2 Dialogue . . . 53
  - 2.2.5 Description of Analyses . . . 54

3 Results: Chance Analysis **56**

4 Results: Predicting Task Performance **59**
4.1 Predicting Task Performance . . . 60
4.2 Learning Effects . . . 65
4.2.1 Uncertainty Elicitation Task . . . 65
4.2.2 Diapix Task . . . 66
4.2.3 CSAR Task . . . 67
4.2.4 Map Task: Completion Time . . . 68
4.2.5 Map Task: Path Deviation . . . 68
4.3 Team Performance . . . 69
4.3.1 Map Task . . . 70
4.3.2 Diapix Task . . . 70
4.3.3 CSAR Task . . . 70
4.3.4 Uncertainty Task . . . 70

5 Results: Mediation Analyses **72**
5.1 Uncertainty Elicitation Task . . . 73
5.2 Map Task . . . 74
5.3 Diapix Task . . . 75
5.4 CSAR Task . . . 76
5.5 LIWC Model Validity . . . 76

6 Results: Correlations between Models **78**

7 Results: Predicting Accuracy vs. Time **82**
7.1 Uncertainty Task Accuracy . . . 83
7.2 Word Count . . . 83

8 Results: Additional Communication Metrics **89**
8.1 Grounding: Length of Installment . . . 89
8.2 Articulation Work: Pronouns . . . 91
8.3 Articulation Work: Swearing and Negative Emotion . . . 91

9 Discussion **94**
9.1 Summary of Results . . . 94
9.2 Common Ground and Task Performance . . . 95
9.2.1 Communication Processes Predict Task Performance . . . 95
9.2.2 Coordination Bests Alignment . . . 99
9.2.3 Which Variant of Coordination . . . 101
9.3 Common Ground: Measurement and Task . . . 103
9.3.1 Bridging Content to Quantitative Measures . . . 103
9.3.2 The Track 2 Dialogue Model . . . 104
9.3.3 Recurrence Analysis of Communication . . . 104
9.3.4 Task Characteristics are Crucial . . . 110
9.4 Applied Relevance . . . 113

9.4.1 Articulation Work 115
9.5 Future Work 116
9.5.1 Strengthening the Method 117
9.5.2 Future Applicability 118

10 Conclusion **120**

References **121**

A Appendix A: Recurrence Metric Chance Analyses **137**

B Appendix B: Complete Recurrence Models **140**

C Appendix C: Detailed Learning Analyses **159**

List of Figures

1.1 Example recurrence plots using the transcript of *Green Eggs and Ham*. The left panel shows an RQA of the transcript with itself. The right panel shows a CRQA of the transcript with a shuffled transcript. 20
1.2 Illustration of the recurrence tests for alignment, coordination, and baseline (adapted from Fusaroli & Tylén). Alignment models were sensitive to patterns transferred between speakers. Coordination models were sensitive to patterns independent of speaker, which included patterns across speakers as illustrated here. Baseline models were sensitive to patterns within one speaker (i.e., self-consistency). (Figure used with permission from John Wiley and Sons; Note: The original figure in Fusaroli & Tylén referred to coordination as interpersonal synergy). 22

2.1 An example map from the HCRC Map task. 36
2.2 Screen shot from the AFRL Uncertainty Elicitation task. This shows a push-to-talk trial with a clear overhead map and a reduced contrast set size (indicated by color-coded street-level views). Building numbers appear in yellow on the overhead map and participants labeled street-level images using the drop-down boxes centered on each row of images. 40
2.3 Two images from the Diapix task that illustrate the differences between the images that the partners were instructed to find through describing the pictures to each other. One partner would get the left image and one partner would receive the right image. For an example difference, the seesaw is colored white in the left image and colored green in the right image. 43
2.4 A wide-angle picture from inside one of the virtual reality facilities used in this task. The screen on the right shows one of the enemy forces. 46
2.5 A screenshot from the CSAR task that illustrates a landmark, in this case, a conventional water tower. 46

2.6 An illustration of how categorical time series were constructed for the alignment analysis, using an example from the word level of the Uncertainty Elicitation task. Speaker A is shown at top and Speaker B is shown at bottom. First, words were given a unique numerical identifier, shown here beneath each word. Then, the silence identifiers -1 and -2 were added to Speaker A's and Speaker B's respective time series. Note that there is no time step for which both Speaker A and Speaker B are silent. 50
2.7 An illustration of how prosodic time series were constructed for the alignment analysis. The X axis in each panel shows the sample position. At top, a 2000-sample pitch trace for each speaker. Silences were identified (shown in gray) and removed from each speaker. At middle, the data were much shorter and were normalized. At bottom, both final time series are shown overlaid on the same axes. Some silence was put back with placeholder values that preserved sequencing. Periods where both speakers were silent were eliminated, as indicated by the shortened time series. 52

4.1 Box plots of dependent measures for each task. Seconds are shown for all plots except Path Deviation. The box indicates the inter-quartile range (Q1-Q3) and the bold line indicates the median value. The whiskers indicate 1.5 times the inter-quartile range and outliers appear as open circles. 61
4.2 Overview of recurrence models prediction of final task completion times in the Uncertainty Elicitation task. (* $p < .05$; ** $p < .01$; *** $p < .001$) 62
4.3 Overview of recurrence models' prediction of final task completion times in the Diapix task. (* $p < .05$; ** $p < .01$; *** $p < .001$) 62
4.4 Overview of recurrence models' prediction of task completion times in the CSAR task. (* $p < .05$; ** $p < .01$; *** $p < .001$) 63
4.5 Overview of recurrence models prediction of time to completion in the Map task. (* $p < .05$; ** $p < .01$; *** $p < .001$) 63
4.6 Overview of recurrence models' prediction of path deviation score in the Map task. Note: the ordinate range differs from the other plots. (* $p < .05$; ** $p < .01$) 64
4.7 Mean team completion times for the Uncertainty Task. Error bars show $\pm$ 1 standard error. 71

6.1 Correlation matrix for the pitch level for each task. Note: correlations with $p > .01$ are omitted. 79
6.2 Correlation matrix for the rhythm level for each task. Note: correlations with $p > .01$ are omitted. 79
6.3 Correlation matrix for the morpheme level for each task. Note: correlations with $p > .01$ are omitted. 80
6.4 Correlation matrix for the word level for each task. Note: correlations with $p > .01$ are omitted. 80
6.5 Correlation matrix for the syntax level for each task. Note: correlations with $p > .01$ are omitted. 81

7.1 Overview of recurrence models prediction of first submission accuracy in the Uncertainty Elicitation task. Note: the ordinate range differs from the other plots. (* $p < .05$; ** $p < .01$) 85
7.2 Recurrence Rate values (RR) as a function of word count. Synthesized data by shortening one trial (black) overlaid on the original data (gray). 87
7.3 Determinism values (DET) as a function of word count. Synthesized data by shortening one trial (black) overlaid on the original data (gray). 87
7.4 Average Line Length values (L) as a function of word count. Synthesized data by shortening one trial (black) overlaid on the original data (gray). . . 88
7.5 Line Entropy values (ENTR) as a function of word count. Synthesized data by shortening one trial (black) overlaid on the original data (gray). 88

List of Tables

1.1 Example excerpt from the classic referential communication task of tangrams illustrating the development of grounded material and the increase in communication efficiency (Brennan, 2000). 7
1.2 Common measures calculated from the recurrence plots. Descriptions of each item can be found in the text. 20
1.3 At top, Clark and Schaefer's (1989) 5 types of evidence of understanding. At bottom, Clark and Schaefer's (1987) states of understanding. 30

2.1 Summary of the differences between corpora. Accuracy′ was an extracted performance metric. 33
2.2 Examples of each type of dialogue act annotated in the Map task corpus. . . 35
2.3 Excerpt from the HCRC Map task corpus. The guide ('g') described the route to the follower ('f'). This excerpt illustrates how differences between the maps reduced the information shared by the partners, which led to additional communication (#13-20). 37
2.4 An excerpt from the AFRL Uncertainty Elicitation task. The dyad began by discussing a street-level picture (#8-9) then looking to the overhead map to label it with a number (#10-13), and repeated this sequence. 41
2.5 An excerpt from the AFRL Diapix task Corpus from a Farm scene. In this example, the background noise was at a high-level and some requests are made to repeat whole utterances (#9) or certain words (#15). 44
2.6 An excerpt from the CSAR task Corpus that illustrates how participants had to change plans due to the enemy forces. 47
2.7 Example part-of-speech (POS) tags in the Penn Treebank format that are output by the Stanford Log-linear Part-Of-Speech Tagger. 49
2.8 Recurrence quantification analysis values used in the continuous data analyses. Smaller sample rates had to be used due to practical limitations. . . . 53

3.1 Table showing p-values of chance analyses where recurrence measured was not significantly different from shuffled time series (shown in bold). 58

4.1 Correlations testing for learning in the Uncertainty Elicitation task. All df = 6. 65

4.2 Correlations testing for learning in the Diapix task. All df = 4. 66
4.3 Correlations testing for learning in the CSAR task. Degrees of freedom (df) varied between teams due to missing data. 68

5.1 Uncertainty task mediation analysis for LIWC—See text for details. (*$p <$ .05, **$p <$ .01, ***$p <$ .001) 74
5.2 Map task completion time mediation analysis for LIWC—See text for details. (*$p <$ .05, **$p <$ .01, ***$p <$ .001) 75
5.3 Map task path deviation mediation analysis for LIWC. Step 3 was not conducted because Step 1 was not successful—See text for details. (*$p <$ .05) . 75
5.4 Diapix task mediation analysis for LIWC. Step 3 was not conducted because Step 1 was not successful—See text for details. (*$p <$ .05) 75
5.5 CSAR task mediation analysis for LIWC. Step 3 was not conducted because Step 1 was not successful—See text for details. 76

7.1 Correlations between word count and task performance—See text for details. (***$p <$ .001) 84
7.2 Summary of word count mediation tests for completion time measures. β and β' are the standardized regression coefficients without and with word count in the model, respectively. (**$p <$ .01, ***$p <$ .001). 85

8.1 Regressions on average length of installment using word-level models. Regression coefficients (and standard error) are shown for each recurrence metric of each model. 90
8.2 Regressions on Swearing. Regression coefficients (and standard error) are shown for each recurrence metric of each model. 92
8.3 Regressions on Negative Emotion. Regression coefficients (and standard error) are shown for each recurrence metric of each model. 93

A.1 Results of chance analyses using shuffled controls, showing *p*-values from two-sided paired t-tests. All Map task tests had *df* = 127. All Uncertainty tests had *df* = 39. 138
A.2 Results of chance analyses using shuffled controls, showing *p*-values from two-sided paired t-tests. The non-significant tests are shown in bold. All Diapix task tests had *df* = 47, except the word-level alignment metrics DET and ENTR had df = 39 because no lines existed in 8 of the shuffled recurrence plots. All CSAR tests had *df* = 119. 139

B.1 Uncertainty task, Alignment Models. 141
B.2 Uncertainty task, Coordination Models. 142
B.3 Uncertainty task, Baseline Models. 143
B.4 Diapix task, Alignment Models. 144
B.5 Diapix task, Coordination Models. 145
B.6 Diapix task, Baseline Models. 146
B.7 Map task Path Deviation, Alignment Models. 147
B.8 Map task Path Deviation, Coordination Models. 148

B.9 Map task Path Deviation, Baseline Models. 149
B.10 Map task Completion Time, Alignment Models. 150
B.11 Map task Completion Time, Coordination Models. 151
B.12 Map task Completion Time, Baseline Models. 152
B.13 CSAR task, Alignment Models. 153
B.14 CSAR task, Coordination Models. 154
B.15 CSAR task, Baseline Models. 155
B.16 Uncertainty task First Submission Accuracy, Alignment Models. 156
B.17 Uncertainty task First Submission Accuracy, Coordination Models. 157
B.18 Uncertainty task First Submission Accuracy, Baseline Models. 158

C.1 Uncertainty Task, Alignment Models after Controlling for Learning. 160
C.2 Uncertainty Task, Coordination Models after Controlling for Learning. . . . 161
C.3 Uncertainty Task, Baseline Models after Controlling for Learning. 162
C.4 Map Task Path Deviation, Alignment Models after Controlling for Learning. 163
C.5 Map Task Path Deviation, Coordination Models after Controlling for Learning. 164
C.6 Map Task Path Deviation, Baseline Models after Controlling for Learning. . 165
C.7 Map Task Completion Time, Alignment Models after Controlling for Learning. 166
C.8 Map Task Completion Time, Coordination Models after Controlling for Learning. 167
C.9 Map Task Completion Time, Baseline Models after Controlling for Learning. 168
C.10 CSAR Task, Alignment Models after Controlling for Learning. 169
C.11 CSAR Task, Coordination Models after Controlling for Learning. 170
C.12 CSAR Task, Baseline Models after Controlling for Learning. 171
C.13 Diapix Task, Alignment Models after Controlling for Learning. 172
C.14 Diapix Task, Coordination Models after Controlling for Learning. 173
C.15 Diapix Task, Baseline Models after Controlling for Learning. 174

Acknowledgments

PhDs don't happen on their own, and my case amplifies this common observation. My unusual combination of affiliations (or "masters" if you will), leads me to be grateful to many people at Wright State University, at Infoscitex Corporation, and at the Air Force Research Laboratory.

First, I would like to thank my advisor Valerie for her impressive example of tenacity and scholarship. She always had ambitious expectations for me and the other students in the lab and we are so much the better for it. I cannot thank you enough for your mentorship and advice, including on this project but also much beyond it. I also want to thank my committee for their generous provision of time and thought: Scott Watamaniuk, Ion Juvina and Sid Horton. This dissertation has benefited greatly from their insight and support. Many other faculty have invested in me and enriched my training and I must express my thanks to them, notably Robert Gilkey and John Flach. Reflecting on my student life come to a close, there are many schoolmates and friends who have traveled with me along this great adventure. To my long-time friends Steve Gabbard, Julio Mateo and Jordan Haggit, it has been fantastic walking alongside each other as we advance our careers (or second careers). To my fast friends, Drew Hampton, Claire Shah, Beth Bullemer, other siblings in the lab, we have sharpened each other so much in such a short time. I'm anxious to see what else we can do in the future.

At Infoscitex Corporation, I had an usual amount of liberty and responsibility that grew me as a principal investigator and entrepreneur. Tom Hughes, Beth Rogers, Mark Axtell, and Kyle Behymer and many others made a big investment in my professional development. They allowed me numerous opportunities to stretch out and try new things, often in high-stakes and fast-paced environments but with great confidence in me.

At the Air Force Research Laboratory, I have had wonderful colleagues from a swath of disciplines that has been essential to my growth as a researcher and collaborator. Brian Simpson, Griffin Romigh, Eric Thompson, Nandini Iyer, and Nia Peters have a terrific lab

atmosphere that I was delighted to be a part of. This project wouldn't have been possible without the flexibility they provided, as well as the dialogue corpora that were collected and shared (and the nice computers didn't hurt either).

For this dissertation project, I must thank Riccardo Fusaroli for his interest in my project and his availability to answer my often long and detailed emails. In addition, while teaching myself recurrence quantification at the onset of this project, I had very helpful discussions with Rick Dale and Jay Holden that saved me a tremendous amount of time and steered me away from some pitfalls.

Outside of my professional life, there have been some tremendous supporters, friends and family, that helped me see this PhD through. My parents Melinda Jenkins and Ed Rothwell have been wonderful advocates for me, being present and thoughtful even from afar and amidst their own busy lives. My friends of the Dayton C.S. Lewis bookclub Tom Bullard, Tim Tuinstra and Todd Bailie have been regular sounding boards for the challenges of balancing work, school, family. They've also provided a welcome and stimulating relief to think about things other than the nearly-obsessive focus that comes with grad school. My friends from Washington Heights Baptist Church and specifically my dear Home Group have encouraged me so much and saw me through some of the most challenging periods of this undertaking.

Most importantly, my wife Jessica has done as much and likely more work than I have for this achievement. She has sacrificed greatly and deserves a large share of this degree. I love you so much!

This project is dedicated to Beatrix and Campbell.

I’m so happy to be a part of your lives and I’m eager to watch them unfold.

Introduction

Language is pervasive in human experience. It is exceedingly difficult to imagine life without it, yet Schaller (2012), in the book *Man Without Words*, describes meeting Ildefonso, a 27-year old who was deaf and pre-lingual. Schaller describes the first time Ildefonso recognized the meaning in the hand gesture symbols and pictorial symbols she was using–the exact moment he first understood language.

> "Suddenly [Ildefonso] sat up, straight and rigid, his head back and his chin pointing forward... My body and arms froze in the mime-and-sign dance that I had played over and over for an eternity. I stood motionless in front of the streaked *cat*, petted beyond recognition for the fiftieth time, and I witnessed Ildefonso's emancipation.
>
> He broke through. He understood. He had forded the same river Helen Keller did at the water pump when she suddenly connected the water rushing over her hand with the word spelled into it. Yes, w-a-t-e-r and c-a-t *mean* something. And the cat-meaning in one head can join the cat-meaning in another's head just by tossing out a *cat*...
>
> He had entered the universe of humanity, discovered the communion of minds. He now knew that he and a cat and the table all had names... Welcome to my world, Ildefonso, I thought to myself. Let me show you all the miracles accomplished with symbols, all the bonds and ties between human beings,

young and old, and even with those dead for centuries." (p. 44-45, emphasis original)

Language provides for the remarkable *joint* project of communing minds. This joint project creates and uses shared meaning, the 'cat-meaning' shared between Schaller and Ildefonso. Shared meaning is also known as common ground and it is developed, maintained and repaired through the process of conversational grounding. More formally, conversational grounding is the process by which interlocutors (the participants in a dialogue) come to understand each other and build up common ground. Conversational grounding is a quagmire for practically-oriented computer engineers, programmers, and application designers involved in human-computer interaction with any type of symbols, but it is particularly apparent when computers attempt to use language.

Philosophers make an important distinction between two types of symbol meaning: extensional semantics and intensional semantics. Extensional semantics refers to the connection between the symbol and the world, also known as the symbol grounding problem. In extensional semantics, the symbol is referencing some *thing* (or, more precisely, the perception of some thing). Robotics is tied to extensional semantics particularly because robots act in the physical world, and roboticists such as R. Mooney have been concerned about the symbol grounding problem for years (e.g., Mooney, 2008; Thomason et al., 2016). Symbol grounding specifies how the robot's representation is mapped to reality.

Outside of robotics, and more often, the symbols are intensionally grounded to other symbols. These could be an aspect of the computer itself, such as a symbolic representation in memory. For instance, the windows operating system shows lists of files in a directory and those labels refer to the data of the file that is stored in memory. The technical problem that concerns us here is when the symbols that need to be related are held by different agents. Conversational grounding is established and agreed upon between two agents as they collaborate as partners. From a practical perspective, errors arise when the partners in collaborative work do not agree on symbol meaning and can not resolve their

misunderstanding.

Consider two examples from aviation mishaps: Turkish Airlines 1951 and Asiana 214. On 25 February, 2009, Turkish Airlines Flight 1951 crashed during its approach to the Amsterdam Schiphol airport. Many factors contributed to this accident, one of them being the meaning of the 'RETARD' mode of the autopilot presented to the pilot on the primary flight display. The pilot was executing a normal approach—bringing the plane in for a landing while utilizing the autopilot for airspeed control (i.e., autothrottle). However, there are two types of RETARD, one for flight level changes and one for flaring to land, and the primary flight display annunciation panel doesn't distinguish between the two (Silva and Hansman, 2015). The pilot's dependence on the autothrottle would have been appropriate for the RETARD for flight level changes but not for the RETARD for flaring to land. The meaning of RETARD for the system was based on the altitude information reaching the autopilot. The autopilot believed the aircraft to be below 27 feet in altitude because the autopilot was receiving and using erroneous altitude data indicating the aircraft height at -8 feet, which disagreed with the altitude data presented to the pilot-flying. The pilot's primary flight display showed a conflicting but correct altitude status, leading to confusion over the situation. Ultimately, misunderstanding about the meaning of RETARD between the pilot and autopilot led to an unrecoverable stall. The aircraft crashed killing 9 and injuring 117 (Dutch Safety Board, 2010).

An accident of similar origin occurred when Asiana Flight 214 crashed on July 3rd, 2013 during approach to San Francisco International Airport. The pilot flying placed the autopilot into a 'HOLD' mode, without fully understanding what HOLD meant for the automatic airspeed control. One notion of HOLD is maintain the current setting, but another notion is similar to when a phone call is placed on HOLD (i.e., stopped). In this circumstance, HOLD meant the latter, which deactivated automatic airspeed control. This led the aircraft to lose altitude and collide with a sea wall, killing 3 and injuring 187 (National Transporation Safety Board, 2014). Both accidents are cases of lexical ambiguity,

where humans and machines did not share the meaning of a word. These stories illustrate that, despite being a philosophical problem, intensional semantics and the conversational grounding problem have important implications for system design.

The problem is by no means restricted to aviation. Computers accomplish and inform actions in a variety of domains, the direction and monitoring of which relies on shared understanding of symbols. People use computers to retrieve weather information from the internet, to compose business contracts, to manage the power grid, etc. Even in the case of personal entertainment, such as browsing videos on YouTube, people need to scan available videos, select them, adjust the volume, and pause during interruption (like the phone ringing), rewind to repeat content or fast forward when bored. While playing video games, the simulated environment poses tasks (find the treasure, kill the bad guys, etc.) and the computer mediates those actions (in the virtual game world).

Clearly, computers are tools for action and the great promise of natural language interaction with computers appears within reach. A recent survey of 1,500+ technology experts on the Internet of Things predicted that speech interfaces will be one of the major advances between now and 2025 (Pew Research Center, 2014). Indeed, the last five years mark an up-turn in language technology known as *spoken dialogue systems*, including: commercial personal digital assistants (e.g., Apple Siri, Google Now, Google Assistant, Amazon Alexa, Microsoft Cortana, Samsung Bixby), in-vehicle infotainment systems (e.g., Apple CarPlay, Google Android Auto, Nuance Dragon Drive, Ford Sync), and open-source tools (e.g., Mycroft.ai, Rasa.ai). Moreover, devices that have a speech interface are now commonplace. Speech-based digital assistants feature prominently on smartphones, which are present in 84% of American households (Olmstead, 2017). Over 33 million "voice-first" devices (e.g., Amazon Echo, Google Home) were expected to be in circulation by the end of 2017 (VoiceLabs.co, 2017).

Among the expected benefits of dialogue systems is the promise of handling complex commands with little to no device-specific training. Popular culture promotes a fantas-

tical vision of eventual capability, apparent in movies such as *Her* (Jonze, 2013) and *Ex Machina* (Garland, 2015). Yet prominent philosophers such as Searle (1990) cast doubt on the available technology for such applications. Common computer responses such as "I'm sorry. I didn't get that" or "I don't understand" are not wholly due to limitations with speech recognition, but rather something far more challenging.

1.1 Human Grounding Process

In exchange between humans, conversational grounding is a collaborative process that has been an enduring topic in human-human communication research (Clark and Wilkes-Gibbs, 1986; Clark and Brennan, 1991; Branigan et al., 2000; Pickering and Garrod, 2004). Grounding benefits collaboration by providing a representation of shared context, supporting immediate feedback of actions, and allowing for incremental progress in conveying intent (Brennan, 1998). Recent findings suggest that grounding and repair are universal aspects of communication that cross language boundaries (Dingemanse et al., 2015).

The human-human communication research on grounding has had a long-time focus on how a speaker produces a description in order to make a definite reference to some item of interest. In a typical so-called referential communication task, partners receive ambiguous visual items, such as tangrams, and must work together to manipulate these items in some way, such as placing the items in a specific order. In the process, the partners must make definite references that uniquely specify which of the items they are currently proposing to manipulate. These initial referential expressions may not be successful and both partners work together to ground them so they agree that they've understood each other. Once grounded, the expressions are reused in subsequent trials and render communication efficient. In the example transcription below (Brennan, 2000 from a corpus collected by Stellmann & Brennan, 1993), two partners order the same set of tangrams multiple times over different trials, illustrating the development of grounded referring expressions, and as

a result their communication becomes more efficient over time. Efficient communication results in efficient task completion.

1.2 Process-Outcome Relationship and Task Characteristics

This link between communication process and task outcomes is critical. Brennan (2000) showed that the low-level phenomenon of grounding processes can have a large impact on outcomes in a referential communication task. In that example, grounding occurred over time and as a result, communication became more efficient. The contribution of grounding processes to team effectiveness is apparent in other laboratory tasks, largely measured by task completion time (Clark and Wilkes-Gibbs, 1986; Clark and Krych, 2004; Fusaroli and Tylén, 2016; Reitter and Moore, 2014). Researchers have argued the importance of the process-outcome link, particularly in regards to how communication processes can have a large influence on team outcomes (e.g., Cooke and Gorman, 2009; Kiekel et al., 2002; Gorman et al., 2004; Svensson and Andersson, 2006; Oser et al., 1991).

Two studies have quantified the relationship between communication and performance. Kiekel et al. (2002) investigated the relationship between communication content and task performance in a team unmanned aerial vehicle control task. The performance measure was a composite of many items, such as, number of mission objectives completed, amount of fuel consumed, time elapsed with alarms. Using Latent Semantic Analysis (LSA), they measured the semantic distance between an entire trial's communications and a 10-trial subset of the data. To generate a performance prediction, the semantic distance from LSA informed a weighted average of the actual performance scores for the 10-trial subset. They were able to explain 39% of the variance in task performance. Though not using the term common ground, Yee et al. (2017) showed a relationship between closing-the-loop commu-

Table 1.1: Example excerpt from the classic referential communication task of tangrams illustrating the development of grounded material and the increase in communication efficiency (Brennan, 2000).

	Trial 1
A	ah boy this one ah boy all right it looks kinda like, on the right top there's a square that looks diagonal
B	uh huh
A	and you have sort of another like rectangle shape, the like a triangle, angled, and on the bottom it's ah I don't know what that is, glass shaped
B	all right I think I got it
A	it's almost like a person kind of in a weird way
B	yeah like like a monk praying or something
A	right yeah good great
B	all right I got it
	Trial 2
B	9 is that monk praying
A	yup
	Trial 3
A	number 3 is the monk
B	ok

nication and task performance, measured as completion time in a team version of the Tower of Hanoi. Their analysis relied on manual annotation of dialogue acts. Closing-the-loop was calculated as the number of dialogue acts that initiated (i.e., question, observations, or commands) or closed (i.e., verbally acknowledged) a new joint project as a proportion of the total number of dialogue acts. A higher amount of closing-the-loop communication was related to better task performance (the standardized regression coefficient reported was .41).

In addition to the importance of predicting performance, there is also an important methodological aspect of the process-outcome relationship. Common ground and understanding are difficult to measure directly, but they can be measured indirectly through performance. For tasks that require common ground and successful dialogue to accomplish, any model that explains variance in task performance is also indirectly capturing shared understanding. When common ground is perturbed, such as from changing the interlocutors in the middle of a tangram task, there are clear changes in task performance (Weber and Camerer, 2003). This inference is foundational to the current research, and to other approaches to testing common ground theories (e.g., Reitter and Moore, 2014; Fusaroli and Tylén, 2016).

However, task characteristics can affect the nature and importance of grounding processes as well as the relationship between common ground and performance. Humans are highly sensitive to the task context in which they are communicating, and they change their behavior. For example, adding time pressure to a referential communication task changes the referential expressions in dialogue (Horton and Keysar, 1996). The communication of helicopter aircrews differs between routine and non-routine periods of flight (Oser et al., 1991). Other research has highlighted how basic speech phenomenon (i.e., disfluencies in speech production) differ across tasks and corpora (Shriberg, 1994; Oviatt, 1995).

The process-outcome relationship will form the basis for discriminating between the different mechanisms that have been proposed for the grounding process. A central theme

of the work presented here will be to vary the task context that is likely to affect communication and grounding processes in particular. Findings that are upheld over a variety of task contexts will provide more compelling evidence of the principal mechanisms than findings that appear only under certain conditions.

1.3 Common Ground Mechanisms

Recent research has focused on how common ground and grounding are accomplished. This question is important both for understanding human-human communication and for advancing how computers use language and improve as participants in dialogue. However, the mechanisms responsible for grounding are unclear (Louwerse et al., 2012; Schober and Brennan, 2003; Horton and Gerrig, 2005). Two separate classes of grounding theories exist, one class that emphasizes the *alignment* and similarity between interlocutors, and one class that emphasizes the *coordination* and complementarity between interlocutors.

1.3.1 Alignment

The increasing alignment of interlocutors over time is the simplest account of common ground and conversational grounding (Pickering and Garrod, 2004). Alignment captures the increasing similarity of the interlocutors through adoption of each other's phonetic, lexical, or syntactic content. Alignment has been also called entrainment, convergence, similarity, and imitation (Branigan et al., 2000). Alignment proponents argue that a simple mechanism can explain most communication phenomena without invoking cognitively intensive models of interlocutors. The key feature of alignment is that it exploits priming, an automatic, covert mechanism in which past experiences influence the likelihood of future contributions. Alignment functions at many levels of lexical complexity, from prosody and phonology to lexical selection and syntax. Alignment at lower levels is thought to

propagate to the semantic level and the situation model of the interlocutors, which forms the basis of a mutual understanding of each other and of the world. Alignment has been suggested as a way to improve language interaction with computers (Branigan et al., 2010; Cowan et al., 2015; Branigan et al., 2003). It is simple to mimic. In addition, humans have been found to align more to computers than they align to other humans (Branigan et al., 2010). For these reasons, alignment has been an influential theory for developers of spoken dialogue systems as a way to facilitate interaction and accomplish grounding (e.g., Buschmeier et al., 2009; Brockmann et al., 2005; Tomko, 2006; Varges, 2006; Janarthanam and Lemon, 2009; DeVault, 2008).

1.3.2 Coordination

The alignment notion of common ground is contrasted with two notions that emphasize coordination. These coordination notions suggest that complementarity and not imitation form the basis of mutual understanding. Perhaps the exemplar of complementarity is the question-answer adjacency pair, which appeared earlier in Yee et al. (2017)'s closing-the-loop communications. A question-answer pair, like every adjacency pair, is made up of two pieces that are adjacent in the dialogue and each piece is contributed by a different speaker (Schegloff and Sacks, 2006). An unanswered question is incomplete, and a question-question pair does not have any resolution, whereas the question-answer fits together and completes the unit. Two different mechanisms of coordination will be discussed here: audience design and interpersonal synergy.

1.3.2.1 Audience Design Coordination

One mechanism will be called *audience design* through this research, though a number of terms have been associated with this position (e.g., collaborative/collective/joint activity,

perspective taking, least collaborative effort). Audience design proposes that speakers[1] employ Theory of Mind and realize that their interlocutors do not necessarily share their knowledge or perspective. Speakers can act on this realization and attempt to strategically *design* their contributions based on what they know (or assume) about their audience's perspective.

Importantly, audience design refers to many things in addition to the different spatial perspective of the audience. The perennial problem in dialogue systems is informing the user what he or she can say, which is why the phrase "*I can do things like...*" is so common at the opening menu (particularly for interactive voice response systems for customer service phone calls). This problem arrises because humans struggle to know the audience and its capabilities—what the dialogue system can do and understand. According to Clark and Marshall (1981), speakers rely on several different sources of information about their audience: physical co-presence, linguistic co-presence, and community co-membership. Physical co-presence means interlocutors share the same environment and this allows interlocutors to form utterances that refer to their shared environment. Linguistic co-presence means that interlocutors employ conversational experience and past interaction to form a body of shared information. Community co-membership means that interlocutors share some group affiliation (e.g., national, regional, generational, professional) that provides a body of common knowledge, such as geography or pop culture, and specific linguistic knowledge such as idioms or technical vocabulary. Audience design also participates in comprehension. The audience can use information about the speaker to constrain message interpretation (Keysar et al., 2000). Disrupting the ability to know your interlocutor's perspective through delaying or eliminating visual information or distorting the visual perspective disrupts collaboration (Gergle et al., 2013, 2004; Clark and Krych, 2004).

Two Tracks of Dialogue Audience design processes are revealed in the distinction

[1] I employ the word "speaker" throughout this document to refer to humans who communicate and to remain consistent with the terminology of speaker and addressee used in psycholinguistics, but this is not to be confused with the sound production devices in the auditory perception community.

between Track 1 and Track 2 dialogue. Track 1 and Track 2 are always present and run in parallel, but serve different purposes. Track 1 dialogue contains the semantic and pragmatic content that is regarded as the primary 'business' of the interaction, while Track 2 dialogue includes the meta-communicative exchanges that serve to manage the conversation and establishes the meaning of contributions (Clark, 1996). Track 2 dialogue addresses how interlocutors handle problems when they occur. Problems can occur for a variety of reasons: people may be poor at knowing someone else's knowledge (Schober and Brennan, 2003) or else don't use it because they are time-pressured, using it would be too complex, or they don't have any knowledge to use (Clark and Wilkes-Gibbs, 1986). In addition to problems related to shared knowledge, problems can occur when the listener is not attending to the conversation, when the speaker misspoke (e.g., commits a word substitution), when the listener misheard, when the speaker detects ambiguity in his/her own utterance, or when the listener misunderstands (Clark, 1994; Clark and Krych, 2004).

The resulting problems are detected because interlocutors rely on evidence of understanding and jointly work to maintain understanding. Speakers produce an utterance and seek displays of understanding or misunderstanding. Listeners are responsible for displaying misunderstanding. Consequently, the absence of such displays is meaningful, providing (weak) evidence of understanding, but listeners may also provide explicit displays of understanding (e.g., so-called backchannel communications *'yeah'* and *'ok'*). Track 2 dialogue is also responsible for management of the topic and the task. Within Track 2 dialogue, interlocutors have two different types of task moves, one that is horizontal, acknowledges a contribution and continues the present task, and one that is vertical and either enters or exits a nested sub-task (Bangerter and Clark, 2003). Moreover, these task moves are accompanied by different but partially overlapping symbols. The horizontal continuations of a task are often signaled by 'uh-huh' and 'yeah' whereas the vertical transitions are signaled by 'ok' and 'all right'. Multiple Track 2 processes contribute to the overall goal of establishing and maintaining shared knowledge.

1.3.2.2 Interpersonal Synergy Coordination

The other proposed mechanism for coordination is *interpersonal synergy* (Dale et al., 2014; Fusaroli et al., 2014; Raczaszek-Leonardi et al., 2014; Raczaszek-Leonardi, 2016; Fusaroli and Tylén, 2016). It is steeped in the conceptualization of cognition based on dynamical systems notions that come from ecological psychology. Ecological notions have been prevalent in human factors and ergonomics, and the use of dynamical systems models is growing continually (for reviews, see Guastello, 2017; Gorman et al., 2017).

Interpersonal synergy is a recent and relatively less examined theory than both alignment and audience design. Like audience design, it emphasizes that common ground is due to complementary behavior rather than imitation. Yet the coordination from interpersonal synergy either does not require intentionality (Fusaroli et al., 2014), or redefines it (see also Gallagher and Miyahara, 2012). In a way, this theory is similar to the simple priming of alignment because of its critique of the cognitively intensive audience design. Rather than coordination coming from intentional design based on information, interpersonal synergy proposes that coordination emerges from the joint systems of interlocutors as characterized by a complex dynamical system. This system is achieving stability in a specific context, and the coordinative interactions become cemented in *interaction routines* that are established by a particular system of interlocutors. The introduction of new interlocutors into an established system of interlocutors disrupts the interaction routines and also the communication (Fusaroli et al., 2014).

1.3.3 Previous Research Investigating Common Ground Models

Previous research has investigated two claims from alignment theory. First is the extent to which alignment phenomena are based on low-level priming. Second, is the extent to which alignment is related to grounding at the semantic and situation level.

To the first point, a number of studies cast doubt on how automatic the alignment phe-

nomenon actually are (recently, Mills, 2014; Healey et al., 2014). One piece of evidence relates to lexical selection and comes from research into speech repairs and speech self-monitoring that was conducted prior to the appearance of alignment theory. If alignment is automatic, then it seems unlikely that speakers would monitor their spontaneous utterances for alignment with the dialogue. Yet Levelt (1983) found that speakers do monitor themselves for alignment and perform what he termed *appropriateness-coherence repairs*. In these instances, the speaker's planned contribution was inconsistent with the prior dialogue, the speaker catches this during or after production, and then repairs the contribution so that it is now consistent. This type of self-repair cannot be accounted for by an automatic process.

In addition to evidence from lexical selection, there is evidence against automatic alignment from syntactic structure. Past alignment research used a scripted-confederate laboratory task to emphasize how interlocutors adopt one another's syntax (Branigan et al., 2000), focusing on one pair of syntactic structures (i.e., double object vs. prepositional object structures). The amount of syntactic alignment is increased in situations where there is also lexical alignment, shown when the interlocutors use the same verb. Research by Weatherholtz et al. (2014) also focused on one pair of syntactic structures and showed that the degree of syntactic alignment is mediated by social factors (which may be intentional but are likely automatic), revealing that the syntactic priming found by Branigan et al. (2000) is actually more complex. In addition, Healey et al. (2014) examined general syntactic alignment across many structures in a conversational setting and attempted to control for the presence of lexical alignment. Findings replicated syntactic priming effects in the presence of lexical alignment, but after controlling for lexical alignment showed that interlocutors did not align on each other's syntactical structure. They actually diverged in syntactic structure. The appearance of syntactic alignment in scripted-confederate tasks may not generalize to spontaneous dialogue, however the measurement of alignment in global fashion may account for the difference in results compared to work on one syntactic

structure.

Fewer investigations address the second point of contention; the relationship between alignment and grounding at either the semantic or situation level. One study investigated alignment in a Wizard-of-Oz simulation, in which a participant played the role of a natural language-enabled robot, and measured alignment in terms of the lexical innovation over time, or new words introduced as the dialogue progressed (Koulouri et al., 2015). They found that alignment was established rather quickly and alignment was predictive of task success. However, when miscommunications or problems in understanding arose, interlocutors solved these by introducing new words and other lexical innovation. The lack of reliance on alignment to resolve problems in understanding suggests that it is not a sufficient model for conversational grounding. Another study, Reitter and Moore (2014), analyzed syntactic priming by attempting to relate syntactical alignment to task performance using a similar rationale to the current research. The authors argued that according to alignment theory, alignment of syntax should contribute to alignment of the situation model, which should in turn contribute to task success. However, they found that short-term syntactic priming was unrelated to performance in the task. In contrast, they found that long-term alignment was related to task performance— long-term alignment was defined as increases in syntactic repetition over a period of minutes, i.e., between the first portion and the last portion of a dialogue compared to the last portion of dialogue from a different block. Their results suggest that an implicit learning mechanism different from priming is at work and that this finding is contrary to the short-term priming put forth by the alignment model of Pickering and Garrod (2004).[2] However, their results did not speak to other types of alignment found in the literature, such as phonetic and lexical alignment and their results do not simultaneously test if a coordination account provides a better ex-

[2] Reitter and Moore (2014) suggest that the alignment model does not specify if priming is a short, priming-based notion or the longer-scale implicit learning that they found. The current research takes a plain reading of priming in Pickering and Garrod (2004) and finds it hard to adopt another reading given the experimental data used to argue for alignment analyze the trial immediately following the prime (e.g., Branigan et al., 2000).

planation of task performance.

1.4 Current Research

The current research attempted to differentiate between the three different accounts of common ground. Analyses applied quantitative recurrence-based models of alignment and coordination, originally developed by Fusaroli and Tylén (2016) (detailed below). An additional analysis extended the original models using statistical mediation to identify if the coordination was audience design or interpersonal synergy. The analyses were performed at multiple levels of linguistic complexity: the prosodic level, the speech/pause level, the morpheme level, the word level, and the syntax level. Previous examinations at the word and syntax levels have only tested grounding theories in isolation. The five levels of analysis also provided an examination into the propagation of alignment across levels.

The analyses were applied to four different team dialogue tasks. The tasks varied in whether or not the performance metric was completion time or accuracy, as well as whether or not it was a symmetric task or an asymmetric dialogue task. Importantly, the tasks and resulting team dialogues are more complicated than the task in Fusaroli and Tylén (2016) and result in richer, longer dialogues with more numerous and diverse content to create problems that stress grounding processes. However, due to the complexity of the tasks, the potential exists that the grounding processes, though relevant to the outcome of low-level tasks, are swamped by more powerful influences on outcome or by the increased variability of the task outcomes.

By investigating a variety of levels of linguistic complexity and four different tasks, the analyses can provide evidence for a general mechanism of communication, rather than just applying to certain situations. Consistent with Fusaroli and Tylén (2016), I hypothesized that the coordination model will have a stronger relationship to performance than the alignment model. Furthermore, I expected that coordination would be the audience design

variant rather than interpersonal synergy. If so, audience design becomes a requirement for human-computer communication.

1.5 Dissertation Organization

The remaining dissertation is organized as follows. First, the general introduction of the relevant literature continues. This includes an overview of the quantitative models that feature prominently throughout this research (Section 1.6). The speech corpus analyses are presented in Chapter 2, where the dialogue tasks are described and the metrics are reviewed. The results are presented in Chapters 3 through 8 and the discussion of results appears in Chapter 9.

1.6 Quantification of Recurrence

I now turn to the quantitative models of grounding processes that will be employed throughout the current research. Separate bodies of research investigate coordination and alignment theory individually, but to the authors' knowledge only one study has attempted to examine both as alternative accounts for the same performance data (Fusaroli and Tylén, 2016). Recent advancements in the analysis of dialogue informed the debate by providing a rich set of metrics for characterizing the relationship between utterances: the non-linear analysis techniques of recurrence quantification analysis (RQA) and cross recurrence quantification analysis (CRQA). RQA and CRQA originated from dynamic systems and were developed to examine recurrence (i.e., repetition) of states in time series data in chaotic systems.

RQA seeks recurrence within one time series (analogous to autocorrelation) and CRQA seeks recurrence between two time series (analogous to cross-correlation). These methods were originally built for continuous data, but they have been adapted for categorical data and used in psychology (Dale and Spivey, 2005), including for lexical analysis (Orsucci

et al., 2013; Fusaroli and Tylén, 2016; Angus et al., 2012a,b), syntactic analysis (Dale and Spivey, 2006), and turn-taking (Gorman et al., 2012). RQA and CRQA can be applied to many different situations as they do not require any data transformations or assumptions of normality. These characteristics make them useful for systems that are non-stationary, that is, systems whose mean or variance changes over time. One requirement is enough samples of the time series to ensure stable metrics, and a rule of thumb is 1,000 data points in the time series (Marwan et al., 2007). Observation times that are too short risk mischaracterizing the system dynamics (Rieke et al., 2004).

Both RQA and CRQA employ a recurrence plot, a graphical depiction of the recurrence in a time series. From the recurrence plot, a number of metrics can be extracted: how much recurrence occurs, how many longer sequences of recurrence occur, the proportion of recurrence that appears in a longer sequence, the average length of recurrence sequences, and the variety in sequence lengths. This in turn allows us to test alternative characterizations of recurrence from the alignment and coordination theories.

1.6.1 Recurrence Analysis Illustration

To illustrate the metrics, consider an analysis of the transcript of *Green Eggs and Ham* by Dr. Seuss. The left panel of Figure 1.1 shows an RQA of the transcript at the word level. The transcript is treated as a time series by using each word as one step in time. The same time series is plotted along the abscissa and ordinate axes. Points appear in the plot where there is recurrence, that is, where a word appears in both time series. Where there are multiple recurring words in a row, a longer structure appears that is parallel to the positive diagonal. These are referred to as *lines*, and one can set a threshold for how many consecutive recurrences are required to count as a line. The right panel of Figure 1.1 shows a CRQA of the transcript forwards and a "shuffled" transcript with the words randomly ordered. Comparison between the RQA of the transcript and the CRQA between the original and the shuffled transcript illustrates how these techniques are sensitive to the structure of

recurrence.

Common measures calculated from the recurrence plots appear in Table 1.2. The recurrence rate (RR) is the proportion of the plot that are points of recurrence. Notice that the recurrence rate was the same between the Forwards-Forwards RQA and the Fowards-Shuffled CRQA. This is because shuffling only changed the order of the words, it did not remove or add any words. The number of lines is a count of how many diagonal line structures are present in the plot, with a line being defined here as 2 or more sequential points along the positive diagonal. Determinism (DET) is the proportion of recurrent points that were part of a line. DET indicates how much recurrence appears as part of a larger sequence. Shuffling the transcript reduced the determinism score dramatically, as the longer sequences of recurrence were disturbed by randomly ordering the words. Similar to the number of lines, determinism decreased greatly when the transcript was shuffled. If the recurrence rate is held constant, the number of lines is highly correlated with the determinism score, as in this example. In other words, with a fixed recurrence rate, the proportion of points that belong to a line (i.e., the determinism) increases as more lines appear in the recurrence plot.

The average line length (L) is calculated as the mean average of the distribution of line lengths in the recurrence plot. Longer average line length corresponds to longer sequences of recurrence. Since these plots were generated by defining a point of recurrence as a single word that appeared in both time series, having a longer average line length means longer sequences of words found in both transcripts. For instance, there are many repetitions of the four word sequence "*I do not like*" found in many lines such as: "*I do not like that*," "*I do not like them here or there*," "*I do not like them anywhere*" and "*I do not like green eggs and ham*." Line entropy (ENTR) is calculated from the distribution of line lengths found in the recurrence plot. More entropy means that there is more variety in the line lengths found (approaching a uniform distribution of line lengths), whereas less entropy means that there

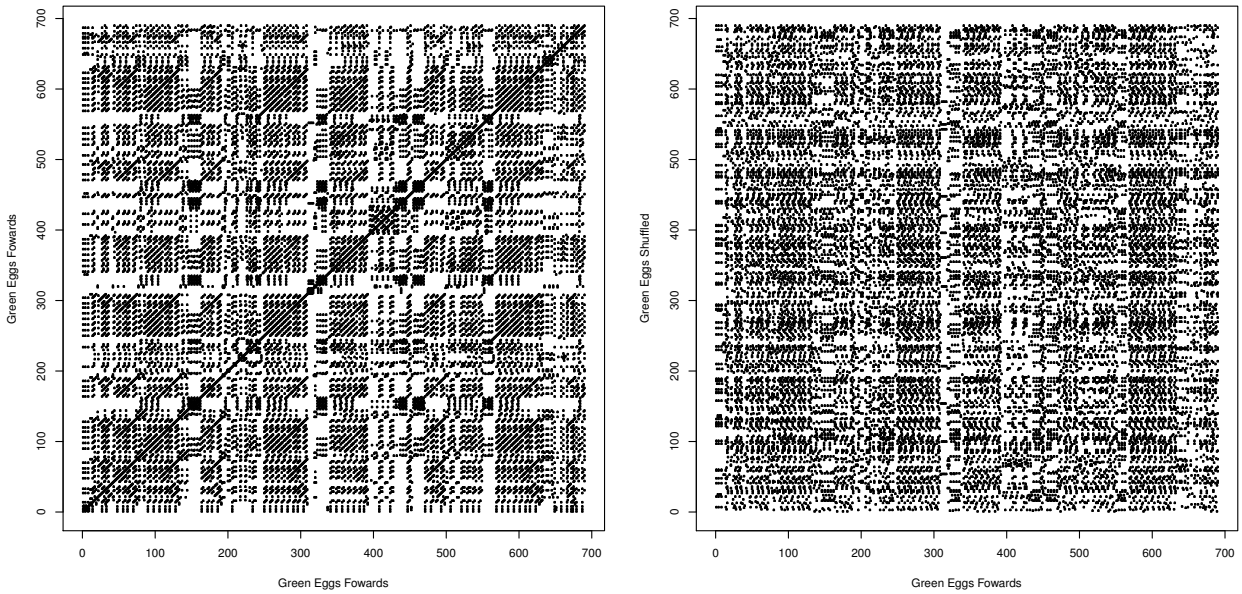

Figure 1.1: Example recurrence plots using the transcript of *Green Eggs and Ham*. The left panel shows an RQA of the transcript with itself. The right panel shows a CRQA of the transcript with a shuffled transcript.

Table 1.2: Common measures calculated from the recurrence plots. Descriptions of each item can be found in the text.

Measure	*Forwards-Forwards*	*Forwards-Scrambled*
Recurrence Rate (RR)	4.40	4.40
Number of Lines (NRLINES)	3189	731
Determinism (DET)	55.12	7.06
Average Line Length (L)	3.62	2.02
Line Entropy (ENTR)	1.54	0.11

are fewer or possibly only one length found.

1.6.2 Grounding Process Models

Fusaroli and Tylén (2016) created models for alignment and coordination using recurrence analyses, and discriminated between these models by their relationship to task performance. Their model of alignment examined recurrence between the two speakers, in which one speaker might adopt or repeat the wording or prosody of the other speaker. Their model of coordination examined recurrence in a speaker-independent manner, in which patterns of interaction between the two speakers might appear multiple times (e.g., adjacency pairs). They compared the two theories for elements at the morpheme, prosodic and speech/pause levels. As a control, they also ran a baseline RQA of each speaker's self-consistency, then used the speaker that had highest rate of recurrence.

The three approaches are illustrated in Figure 1.2. The contents of the time series are the same in all three exchanges, but different patterns are outlined to reflect the different recurrence for which these models are sensitive. In the alignment model, patterns that are transferred from one speaker to the other will be detected, such as 'XYY' from speaker A to speaker B (patterns that go from B to A will also be detected though not shown in the schematic). In the coordination model, patterns that appear independent of which speaker contributes will be detected. This includes patterns across speakers. For instance, the pattern 'YXZXY' occurs between A and B and later B and A. In the baseline model, recurrence of patterns within A and patterns within B were tested separately (i.e., self-recurrence of A and self-recurrence of B) to represent a lack of adaptation to the interlocutor. Whichever speaker had the higher self-recurrence rate was used for analysis. (The coordination model contains the alignment model and baseline model points of recurrence but the resulting recurrence metrics (e.g., RR, DET, L and ENTR) are not necessarily related due to the non-linearity of the analysis.)

The dialogue in Fusaroli and Tylén (2016) resulted from two participants perform-

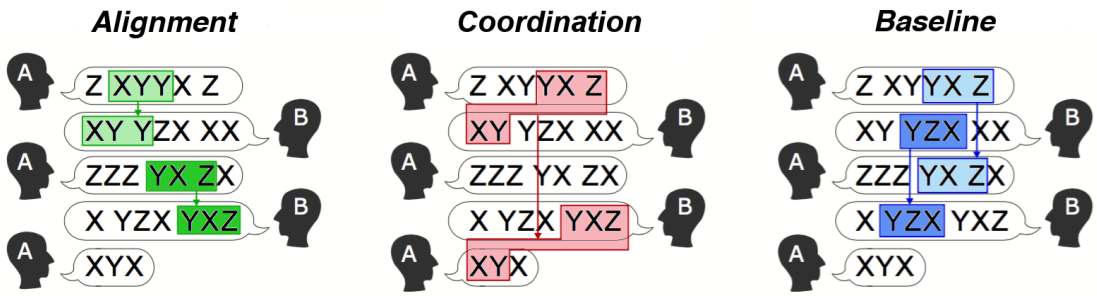

Figure 1.2: Illustration of the recurrence tests for alignment, coordination, and baseline (adapted from Fusaroli & Tylén). Alignment models were sensitive to patterns transferred between speakers. Coordination models were sensitive to patterns independent of speaker, which included patterns across speakers as illustrated here. Baseline models were sensitive to patterns within one speaker (i.e., self-consistency). (Figure used with permission from John Wiley and Sons; Note: The original figure in Fusaroli & Tylén referred to coordination as interpersonal synergy).

ing a two-interval oddball detection task based on the visual contrast of the oddball. Each participant made an independent response on whether the oddball appeared in the first interval or the second interval. When their responses agreed, they automatically proceeded to the next trial without discussion. When their responses disagreed, they discussed their responses and came to a collaborative judgment. Afterwards, the researchers calculated contrast sensitivity for each individual and the joint/collaborative sensitivity. Then a collaborative benefit was computed as the ratio between joint sensitivity and the highest individual's sensitivity. Ratio values greater than 1 indicated a benefit from the joint decision. The recurrence metrics of average line length (L) and entropy (ENTR) were calculated according to each theory. Then, these recurrence metrics were used as predictors in regression models with collaborative benefit as the outcome, thereby relating grounding mechanisms to the task performance benefits for a collaborative laboratory task.

The study found that *both* the alignment and coordination models provided dialogue recurrence metrics that were related to performance in the task, whereas the baseline model was not related. However, the coordination models were better predictors of performance than the alignment models for the lexical level (*Adjusted* R^2, $AdjR^2$= 0.45 vs. $AdjR^2$= 0.25)

and the speech/pause level ($AdjR^2$= 0.31 vs. $AdjR^2$= -0.12). The two models were similar for the prosodic level ($AdjR^2$= 0.31 for both).

1.7 Extending Recurrence Models

The recurrence models from Fusaroli and Tylén (2016) offer promise for testing both alignment and coordination models on the same dialogues. The current research extended their findings and methods through: 1) increasing dialogue complexity, 2) additional levels of analysis, 3) investigating task symmetry, 4) investigating the construct accessed by the coordination recurrence model, and 5) adding additional recurrence metrics as predictors.

Both the nature of the dialogues and the team task in the Fusaroli and Tylén (2016) study are relatively constrained and low-level (the discussion centered on whether or not to select the first or second interval). The tasks selected for the current research promoted complicated dialogues and also had specific features that stressed conversational grounding. For instance, one task had a varying background noise level applied to one or both partners. This challenged the partners' ability to understand each other and promoted compensatory changes in their dialogue. Another task manipulated whether or not the audio was open microphone or push-to-talk, which could affect the dynamics of interactions, such as timing and frequency of backchannel communications. It is unknown if the recurrence models will have the same success on more complex dialogues in richer task environments.

Similarly, the support shown for coordination may or may not be replicated in higher-level tasks or at additional levels of analysis (e.g., the word level or the syntax level). Evidence for alignment has been found for words and syntax (Branigan et al., 2010; Cowan et al., 2015; Branigan et al., 2003), but these levels were not tested by Fusaroli and Tylén (2016). Perhaps the research conclusions will change when these levels are incorporated, however, if coordination is still supported at these levels it will be strong evidence for its adoption. Each dialogue corpus uses a different task with the purpose of replicating

findings in a task independent manner in the same fashion as replication of findings across the various levels of analysis (syntactic, word, prosodic) provides additional evidence of the primary grounding mechanism.

The recurrence metrics are arguably opaque, and the connection between average line length and H. Clark's work on grounding, for example, is not readily apparent. Therefore, additional analyses using statistical mediation examined the grounding mechanism to which the recurrence metrics respond.

These analyses tested if the coordination recurrence model mediates a relationship between Track 2 dialogue features and performance. In this fashion, the coordination variant can be identified as audience design or interpersonal synergy.

Another characteristic that appears across the tasks selected for the corpus analysis is the extent to which the dialogue roles are symmetric or asymmetric. Fusaroli and Tylén (2016) tested a symmetric task only. In a symmetric task there are no defined roles and specifically no "answers" held by one participant that had to be communicated to the other. As a result, the conversational dynamics in symmetric tasks are flexible and negotiable. Partners can dynamically switch roles or avoid set roles all together. Asymmetric tasks, which are more common in the literature, have defined roles for each participant and these roles greatly influence the communication content and dynamics. One role is a *director* or *guide* and has the "answer" (e.g., a route drawn on a map, an example model to build, a particular arrangement of objects), which needs to be communicated to the other participant who is the *matcher* or *follower*. As a result, the different roles affects the dialogue that takes place, which can be clearly seen by analyzing the dialogue acts each interlocutor makes (for one analysis, see Lickley, 2001). The director typically does a lot of describing or instructing and the matcher does a lot of acknowledging and checking understanding. The current work tested if there are consistent findings between symmetric tasks and asymmetric tasks and specifically if coordination is superior to alignment in both circumstances.

1.7.1 Examining Validity

Just as the type of coordination needs further specification, these recurrence models are relatively new and unstudied. Additional investigation into their construct validity was merited, to test that they measure what they purport to measure. For example, Fusaroli and Tylén (2016) found that prosodic-level alignment and coordination models accounted for the same amount of variance in performance. Was the cause of this similarity a lack of differentiation between the models at this level of analysis? What was the relationship between the recurrence metrics that were used as predictors? Analysis into the relationships between the models will inform the extent to which they access the different alignment and coordination constructs. Strong relationships between the alignment model and the coordination, for instance, would indicate problems with construct validity. Construct validity also was established through measuring the relationship to previously established measures of grounding, described next.

1.8 Informing Dialogue Systems

As suggested above, interlocutors are very sensitive to their context and therefore grounding processes can be moderated by task features. In addition, task features are important to understand when trying to make recommendations for dialogue system developers in various domains and circumstances. The previous literature on common ground, including Fusaroli and Tylén (2016), has used simplistic tasks with no or minimal demands inherent in cooperative work environments. One task in the current research was selected specifically to investigate connections between grounding and the construct of task management, a critical aspect of joint activity that resides in a separate body of literature outside of dialogue and converges with the construct of Track 2 dialogue.

Prior work on task features that moderate common ground is rare (Horton and Keysar, 1996; Brennan and Schober, 2001). Some models of grounding leave them out or assume

they are static (Bunt et al., 2007). Other work includes various payoffs in the task but does not analyze how grounding phenomenon may change (Gravano, 2009), performing analyses that aggregate over high and low payoff trials and ignoring potential differences in behavior associated with differences in the reward structure within the task. When research has recognized task features on occasion, they are uni-dimensional features and notions of multi-dimensional features are ignored. The lack of such insight has led to dialogue system developers historically ignoring moderating influences. Restricted application domains such as shopping and navigation fail to reveal the limitations of these underlying assumptions. But as the scope of artificial intelligence grows and natural language processing technologies become more integrated into work practices, their applications will not be limited to the subset of activities with overly simplified grounding processes. The current research will begin to address these moderating influences through one of the selected tasks and hopefully promote interest in such research.

1.8.1 Task Management and Articulation

Past work by Traum and Dillenbourg (1996) displays an exquisite sensitivity to task issues that are central to joint activity, while highlighting some important challenges. They extended an earlier formalization that attempted to predict the contributions of individual communications to task completion while recognizing the difficulty in calculating payoffs and costs for individual communicative intentions. Even when task payoffs and costs are known, there is a great deal of uncertainty about how valuable a specific utterance is at the task level, which complicates any attempt to calculate that value. In their discussion, the authors described that participants' grounding often took the topic of task management, such as problem solving strategy, decomposition of who does what and when. Its unquestionable that task management impacts the team's task performance yet similarly, the costs and payoffs for grounding this task management material are difficult to calculate.

Task management, also known as *articulation work* (Strauss, 1985), takes a promi-

nent status in other research areas focusing on teamwork and team performance, such as computer supported cooperative work (CSCW). Notable researchers Schmidt and Bannon (1992) and Malone and Crowston (1990) have argued that any cooperative work has articulation[3] as an integral part. Articulation work bears a relation to Track 2 dialogue that was discussed in Section 1.3. Where Track 2 dialogue is dialogue about the communication itself and works to clarify the contributions and intentions of the interlocutors, articulation work is discussion and negotiation about the task itself that serves a number of purposes: define or refine the goals of the team, to perform functional decomposition and divvy up responsibilities, to discuss sequencing and temporal coordination, or to highlight or clarify *functional dependencies* that teammates have promised to uphold (Rothwell and Shalin, 2017). Through articulation work, teams "*divide, allocate, coordinate, schedule, mesh, interrelate, etc their individual activities*" (Schmidt and Bannon, 1992 p. 14). A central element of cooperative work is to discuss task management and ground how and to what extent team members are relying on each other (Rothwell and Shalin, 2017). Articulation also plays an ongoing role by monitoring and tweaking the cooperative dependencies during the joint activity (Schmidt and Bannon, 1992).

As an example, consider an investigation of articulation work in teams controlling the London Underground (Heath and Luff, 1992). They highlight the importance of self-talk for working through problems, constraints and the resulting schedule changes. Self-talk is not directed at a particular team member, and does not expect a reply, however it performs a crucial function by updating the team to changes in the situation and the schedule. Teammates are dependent on these updates yet time is rarely available for explicit updates, and frequently the speaker is too engaged in managing the task at hand for a conversation about the updates. Many of the classic communication tasks do not provide opportunity for articulation work and therefore do not reveal these processes and their effect on performance. Articulation work is a broad concept that incorporates many types of team

[3]The CSCW construct of articulation work is not to be confused with articulation as an aspect of speech production mechanisms.

behaviors. As a starting place, the current research incorporated one task that elicited a subset of articulation work phenomena.

1.8.2 Clarification Dialogues in Dialogue Systems

In considering how to characterize common ground for dialogue system development, *clarification dialogues* have been the primary means for spoken dialogue systems to engage in grounding. There are multiple possible sources of problems for natural language understanding that could prompt a clarification: the speaker misspoke or had a disfluency, speech recognition performed one or more word errors, and the utterance was complex and perhaps ambiguous. Human speech analysis challenges automated natural language comprehension (Shriberg, 1994, 2005) because human speech production has disfluencies: repetitions, repairs, filled and unfilled pauses. State-of-the-art attempts at identifying and removing disfluencies have shown great potential (Johnson and Charniak, 2004; Honnibal and Johnson, 2014), but these capabilities are not widely used or widely available. Many researchers appear to downplay problems caused by disfluencies, partially because of larger challenges arising from speech recognition accuracy.

The majority of research has focused on overcoming speech recognition errors. Gabsdil (2003) suggests that systems fall somewhere on a spectrum of clarification strategies between a cautious grounding strategy and an optimistic grounding strategy. The cautious grounding strategy attempts to explicitly confirm all information that the user provides to the system and the optimistic grounding strategy attempts to interpret all user input without clarification. Many systems just adopt one strategy to use universally, though it is possible for a system to employ multiple strategies and decide on a case-by-case basis what strategy should be used, which follows human behavior more closely. Gabsdil suggests that the decision about which clarification strategy to use could be based on a speech recognition confidence metric, and makes no suggestion that different tasks or subtasks of the joint activity may merit different grounding. Similarly, other researchers focus on clarification

of speech recognition errors (Skantze, 2005; Koulouri and Lauria, 2009), leaving out other sources of miscommunication.

This work on clarification dialogues for speech recognition errors has focused on modeling human processes. The findings draw attention to the complexity of clarification—people use a wide variety of clarification methods across different settings and circumstances (Gabsdil, 2003; Skantze, 2005; Passonneau et al., 2012). In the face of these findings, alignment struggles to account for the range of behaviors observed as well as coordination does.

Turning to the smaller body of research on resolving ambiguous statements resulting for example from anaphora and deixis, Allen et al. (1995) and Traum (1994) suggest that utterances are speech acts and successful dialogue systems will identify the intended action. So, much of their work (Heeman and Allen, 1998) and the work of their colleagues (Hirst et al., 1994) is focused on resolving ambiguity in the actions implicated in utterances. Traum's work has been developed into degrees of grounding (Roque and Traum, 2008) and a concept of a common ground unit (Visser, 2011), to manage the dialogue following Clark and Schaefer (1987, 1989)'s types of evidence and states of understanding (Table 1.3). However, these approaches and specifically the computational representations they use are rare with the rise and success of deep learning.

This section concludes by reiterating one of the challenges put forth in Ward and Devault (2015): Dialogue system designers need to engage with social scientists and vice versa. Specifically, they call out that the contributions from social-science research often lack descriptions of behavior that are specific enough for use in developing dialogue systems. Recurrence analyses and dynamical systems analyses are particularly lacking in this regard. The current research attempts to accept their challenge and make progress towards thorough and specific descriptions of the phenomena.

Table 1.3: At top, Clark and Schaefer's (1989) 5 types of evidence of understanding. At bottom, Clark and Schaefer's (1987) states of understanding.

Type	*Description*
Continued Attention	B shows he is continuing to attend and therefore remains satisfied with A's presentation
Next Relevant Contribution	B starts in on the next contribution that would be relevant
Acknowledgment	B nods or says "uh huh," "yeah," or the like
Demonstration	B demonstrates all or part of what he has understood A to mean
Display	B displays verbatim all or part of A's presentation
State 0	B didn't notice that A uttered any *u'*
State 1	B noticed that A uttered some *u'* (but wasn't in State 2)
State 2	B correctly heard *u'* (but wasn't in State 3)
State 3	B understood what A meant by *u'*

1.8.3 Measures of Grounding

One approach to getting specific and useable descriptions of phenomena is through using multiple measures of common ground that have been previously established. The variety of measures will also capture the potentially diverse variety of ways in which conversational grounding processes are influenced by changes in task characteristics. The current research used two established features of the grounding process that are readily computed: the length of installments of a contribution (Clark and Krych, 2004; Brennan, 2004, 1998; Clark and Schaefer, 1989), and the use of pronouns (Khawaja et al., 2012, 2014). The length of installments can be operationalized by measuring the turn length, i.e., the number of words per turn, and the turn rate (the number of turns per unit time). Installment length reduces when speakers need to introduce a complex or important contribution, as they break it into smaller pieces that can each be grounded individually. The pronouns speakers use convey a great deal of information about how they are conceptualizing the team and in particular the dependency relationships in the team that are relevant to articulation work. I expected to see increases in plural pronouns as the demand for articulation work increases.

1.9 Summary of Hypotheses

The research questions of the current work rely on one foundation, that communication processes can predict task performance. I first expected that the models of communication would be related to task performance in the complex tasks used in the current research. The primary research questions of the current research investigated whether the alignment theory or coordination theory accounts for common ground and which variant of coordination was present. I expected that coordination would outperform alignment and that this finding would be consistent across the four separate tasks tested (across the differences in task performance metric and task symmetry). I also expected that coordination would be the audience design variant, shown by a relationship to Track 2 dialogue.

The remaining analyses provided further investigation into these nascent recurrence-based models of common ground processes. One portion of the analyses investigated the construct validity by testing the differences between the models and possible relationships between models. I expected that the alignment and coordination models would be related at the prosodic level of analysis, but not at other levels of analysis. Another portion of the analyses investigated the connections between the recurrence models and previously established measures of grounding as well as demand for task management activities. For both the established measures of grounding and task management, I expected that these would be relevant to task performance in the currently examined tasks. And furthermore, these measures would be related to the coordination recurrence model.

Method

This section describes the methods and tasks used throughout this research, detailing first the corpora, and second the analysis method. To examine task independence, results from four different tasks were compared. All of these tasks provided previously collected *corpora* for other research interests. The tasks have some similar characteristics relevant to the current research. All tasks were dyadic (2-person) tasks that used verbal communication to accomplish the task goals. In all cases, the verbal communications were spontaneous and unstructured. No instructions were given as to what to say or how to best complete the tasks. All the tasks have audio recordings and either existing orthographic transcriptions of the communications or orthographic transcriptions were created for the current research. All the tasks have completion time metrics. One task had an existing accuracy metric and an accuracy metric was extracted for another task. Moreover, these tasks are more diverse conceptually than those in Fusaroli and Tylén (2016), so the dialogues in these corpora have a larger vocabulary. The primary difference between these tasks is symmetry: there are three symmetric tasks and one asymmetric task.

2.1 Materials

Four tasks provided speech corpora: the Human Communication Research Center's (HCRC) Map task, the Air Force Research Laboratory's (AFRL) Uncertainty Elicitation task, the

Table 2.1: Summary of the differences between corpora. Accuracy′ was an extracted performance metric.

Task	# Speakers	Symmetry	Performance Metric
Map Task	2	Asymmetric	Accuracy & Completion Time
Uncertainty Elicitation	2	Symmetric	Accuracy′ & Completion Time
Diapix Task	2	Symmetric	Completion Time
CSAR Task	2	Symmetric	Completion Time

AFRL Diapix task, and the AFRL Combat Search and Rescue (CSAR) task.[1]

2.1.1 Map Task Corpus

The Map task is a team dialogue task from the Human Communication Research Centre. Originally designed to study pronunciation and intonation (Anderson et al., 1991), it has become a classic task dialogue corpus for multiple purposes (e.g., Reitter and Moore, 2014; Branigan et al., 1999; Lickley, 2001).

Participants The study included 64 participants. Some participants knew each other while others did not. Each participant served in two different dyads during the experiment, for a total of 64 different teams. Experimenters assigned one member of the dyad as a guide and the other as a follower.

Task Description In the Map task, two people referred to paper maps containing a variety of labeled landmarks. Participants could only view their own map. The guide's map had a route drawn on it (an example is shown in Figure 2.1). The follower's map had the start marked, but no route. The goal was for the follower to draw the route on his/her map from the guide's descriptions of his/her map. Though a number of landmarks were the same on both maps, differences between the maps complicated the task and instigated communica-

[1]The Battlespace Acoustics branch of the 711 Human Performance Wing collected data for all AFRL tasks.

tion. Some landmarks appeared on both maps but had different labels. Other landmarks were missing from the follower's map or duplicated on the follower's map. These differences perturbed the mutual information held by the partners to perhaps reveal how these differences in information are detected and resolved (i.e., conversational grounding). The 16 pairs of maps used throughout this corpus had different routes and different combinations of landmarks.

Measures Accuracy on the map task was path deviation, calculated by measuring the deviation between the route on the guide's map and the path the follower drew on his map. As an error score, a small deviation corresponded to good performance. Though participants were not instructed to speed task completion, the task completion time may determine the presence of speed/accuracy trade-offs.

In addition, the Map task corpus had annotations of dialogue acts, also referred to as conversational moves (Carletta et al., 1996, 1997). The types of moves fell into either: initiate, response, or ready (Table 2.2). Initiate moves 'introduce a new discourse purpose into the dialogue.' Response moves are in reply to initiate moves and serve to fulfill the discourse purpose that was introduced. Ready moves occurred between dialogue games and functioned to coordinate the beginning of the next game. Four coders rated each utterance with good reliability (K = .83; Carletta et al., 1997).

Task Conditions The experiment employed two manipulations: presence of eye contact and the familiarity of the speakers.

Eye Contact The teams were randomly assigned into one of two groups, a group able to make eye contact or a group that had a barrier separating them and blocking eye contact.

Familiarity The partners were either familiar or unfamiliar with each other. Familiarity may have influenced the dialogue collected through changing the amount of previously established information that can be relied upon.

Table 2.2: Examples of each type of dialogue act annotated in the Map task corpus.

Initiate Moves	**Example**
Instruct	*"go to the right about an inch"*
Explain	*"your mountain must be different from mine"*
Align	*"do you know what I mean?"*
Check	*"okay, up to the top of the stile?"*
Query-YN	*"and do you see a collapsed shelter?"*
Query-W	*"where are you now again?"*
Response Moves	**Example**
Acknowledgement	*"okay"*
Clarify	*"just sort of straight left"*
Reply-Y	*"yeah"*
Reply-N	*"sorry no"*
Reply-W	*"the bottom of it"*
Ready	*"right"*

The 64 participants each completed 4 trials varying in partner familiarity and role (guide or follower). There were a total of 128 trials in this corpus. An excerpt appears in Table 2.3.

2.1.2 Uncertainty Elicitation Task Corpus

The Uncertainty Elicitation task is a team dialogue task designed and collected at the Air Force Research Laboratory. The task was originally designed to elicit spoken uncertainty for the purpose of building computer models to detect and synthesize uncertainty in spontaneous speech (Romigh et al., 2016). It was inspired by the different spatial orientations that Air Force operators have to overcome, such as when people on the ground are coordinating with people in the air.

Participants Ten participants were organized into 5 teams of 2 people. Throughout the task, the teams remained intact.

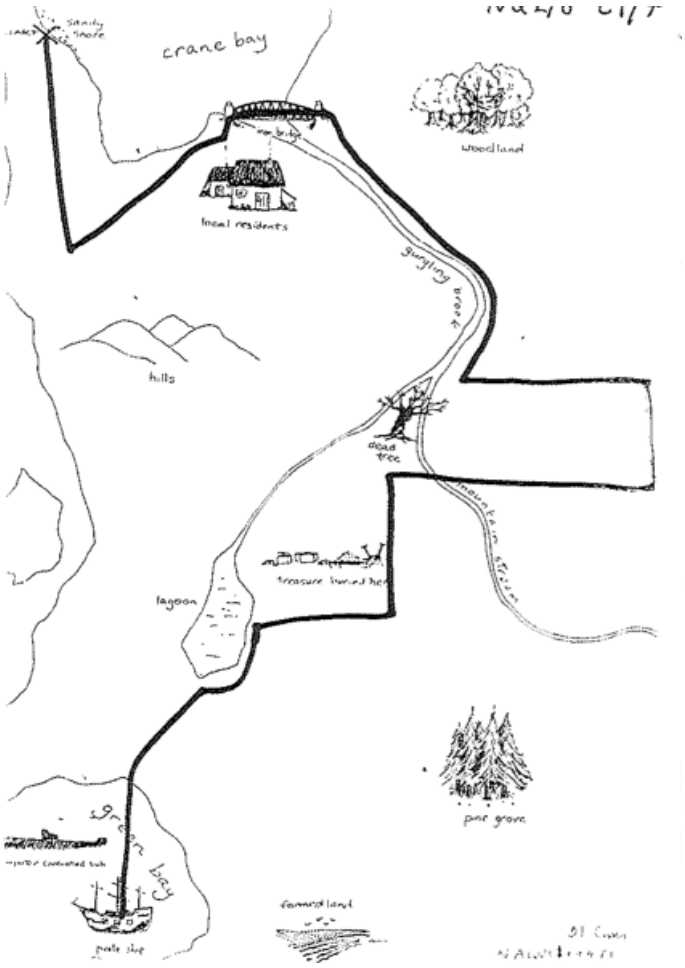

Figure 2.1: An example map from the HCRC Map task.

Table 2.3: Excerpt from the HCRC Map task corpus. The guide ('g') described the route to the follower ('f'). This excerpt illustrates how differences between the maps reduced the information shared by the partners, which led to additional communication (#13-20).

#	Speaker	Transcript	Move
9	g	okay	Align
10	g	and you're going to go down and then you're going to do a "u" shape	Instruct
11	f	uh-huh	Acknowledge
12	g	and we're going to come up and we're going to have the old	Instruct
13	f	eh– ehm are we just going are we going to go below the picket fence	Query_YN
14	g	below what	Query_W
15	f	the picket picket fence	Reply_W
16	g	picket fence	Acknowledge
17	g	I don't have one so I would say	Explain
18	g	whereabouts is the picket fence	Query_W
19	f	picket fence is below the mill wheel which is below the caravan park	Reply_W
20	g	right okay well	Ready
21	g	you want to have the old mill on your right-hand side so if that fence is below the old mill you want to keep that on your right-hand side	Reply_W
22	f	okay	Align

Task Description This task had a referential communication component and a collaborative deduction component. Partners had a shared, labeled, overhead perspective as well as many unlabeled pictures of buildings from street-level perspectives. Participants worked together to label the street-level pictures by identifying the corresponding location of each street-level image on the labeled overhead perspective.

Participants sat in separate rooms with a computer display in each room (Figure 2.2). Google Maps provided street-level pictures and overhead (i.e., satellite) pictures with all Google labels removed. The participants wore headsets and could communicate with each other over a recorded voice loop. Each building appeared in 4 different street-level pictures, each from a different perspective. Each participant had only two of these pictures, prompting referential communication. The participants did not share any street-level views and they had to determine that they were discussing the same building. In addition, the street-level images on each participants' display were in a randomized order that prevented referring to a building by its position on screen.

The overhead map was the same for both participants and had 12 numbered buildings (1-12). The participants had street-level pictures of only 6 of those buildings. Collaborative deduction arose from the need to combine information from the different street-level perspectives to identify the location on the overhead perspective. For example, a picture from the front might show a sidewalk and a different picture from the side might show a swimming pool and patio, and combining these features identifies the building on the overhead map. However, all overhead images were taken at a different (but unknown) time than the street-level images, resulting in differences in the environment (e.g., tree foliage, car placement, and in some instance roof colors), further complicating the task. It was also possible (although rare) that street-level images were not taken at the same time, generally because a corner house was imaged from two different streets at different times.

Participants labeled the street-level views with numbers from the overhead map. The trial ended when both participants had labeled all of the buildings correctly. When partic-

ipants submitted their answers, they received correct/incorrect feedback for each building. The feedback was specific to the individual participant so if the partners failed to ground their discussion *with each other* and mistakenly put down different numbers for the same building, one partner may be correct and the other incorrect.

The task had a repeated measures design. Each of the 5 teams completed 8 trials with different stimuli for a total of 40 trials total in this corpus.

Measures The primary performance metric was completion time after successfully labeling all buildings, with shorter times indicating better performance. Because each team submitted their answers multiple times during each trial, the first submission was extracted as an accuracy metric with a corresponding first submission time. This feature led to two separate analyses of this corpus: one that used the final completion time with perfect accuracy, and one that used the first submission time where accuracy varied.

Task Conditions The 2 x 2 x 2 repeated-measures design manipulated communication channel type, overhead map clarity, and set size.

Communication Channel Type On half of trials, the channel was full duplex communication and on the other half of trials it was half duplex communication. Under full duplex communication both participants could send and receive simultaneously (an open telephone line) and under half duplex communication only one person could send at a time (a push-to-talk radio). Channel voice quality was constant between the two communication channels, though it often differs between real-world telephone and radio. This manipulation may have influenced the dialogue dynamics, such as timing and frequency of backchannel acknowledgements.

Overhead Map Clarity The overhead map was clear on half of the trials and blurry on the other half (i.e., low-pass filtered to reduce details), which could make referential expressions more challenging by complicating how partners identify and name features of the environment.

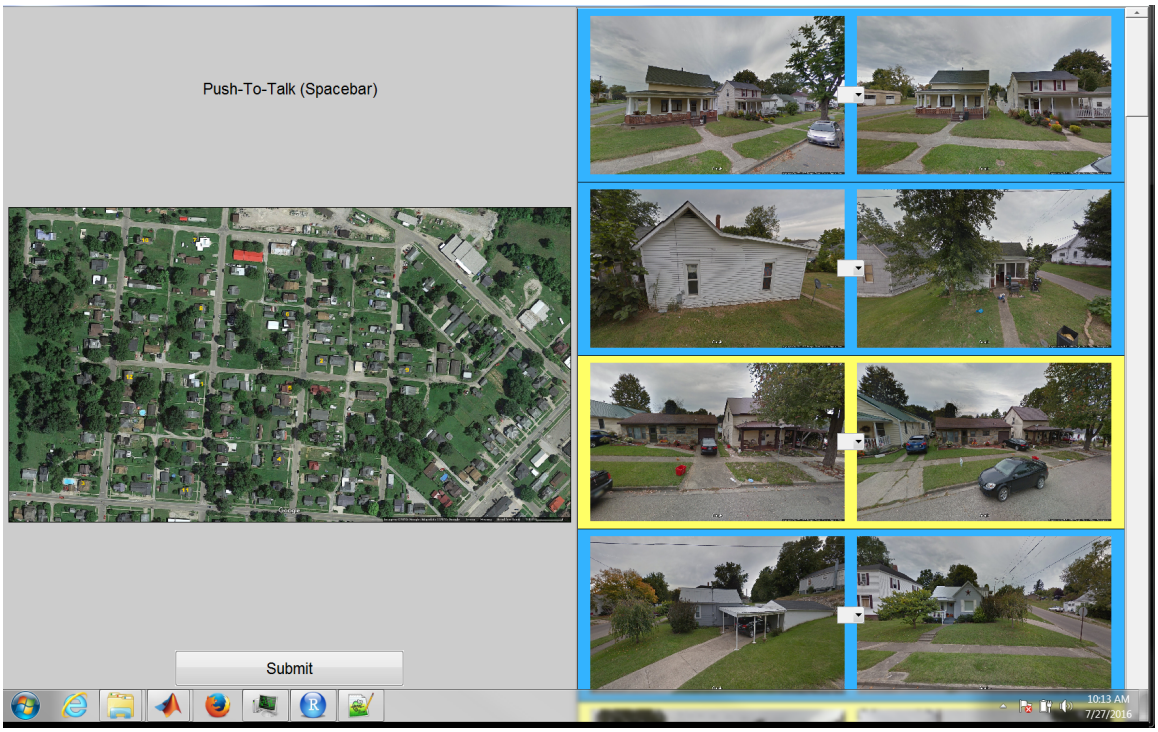

Figure 2.2: Screen shot from the AFRL Uncertainty Elicitation task. This shows a push-to-talk trial with a clear overhead map and a reduced contrast set size (indicated by color-coded street-level views). Building numbers appear in yellow on the overhead map and participants labeled street-level images using the drop-down boxes centered on each row of images.

Set Size On half of the trials, the partners street-level views of the buildings were all the same color so only the referential expressions could be used to distinguish them. On the other half of trials, the partners' street-level views of the buildings were color-coded blue or yellow in order to reduce the set from 6 buildings to 3 buildings (as shown in Figure 2.2), which simplifies grounding by having few alternatives to consider. A transcript excerpt appears in Table 2.4.

2.1.3 Diapix Task Corpus

Data from the Diapix task were collected by the Air Force Research Laboratory. The task was originally developed to study the effect of disparate language backgrounds (i.e., differ-

Table 2.4: An excerpt from the AFRL Uncertainty Elicitation task. The dyad began by discussing a street-level picture (#8-9) then looking to the overhead map to label it with a number (#10-13), and repeated this sequence.

#	Speaker	Transcript
6	B	like let me describe them to you this time maybe that'll help
7	A	alright
8	B	alright so I have a picture that it's like the house is surrounded by trees and bushes do you see it
9	A	yeah
10	B	okay let's try to find that one
11	B	I want to say it's eleven just by the picture of it
12	A	yeah I'm gonna hit eleven too
13	B	okay
14	B	um there's another one it's a small one-story house uh garage is separated
15	B	it's like really small and tiny
16	B	it's the yard is really big there's a tree in the front yard
17	A	I see it and there's like a stop sign
18	B	got it yeah I think it's at a corner
19	A	yeah I see the corner
20	B	okay
21	A	um
22	B	it's really small and very open
23	B	so I'm guessing it's either
24	A	it could be seven
25	B	seven
26	A	I don't know
27	B	I don't think it's seven it's either one
28	A	I– it can't be one because there's a pool in the back of one
29	B	true true true

ent L1) on communication (Van Engen et al., 2010), and extended to study communication challenges with a hearing impairment (Baker and Hazan, 2011). AFRL studied how speakers adapt their speech when their interlocutor is in a mis-matched acoustic environment (Iyer et al., 2016).

Participants There were 16 participants organized into 8 teams of 2 people. Throughout the corpus, the teams remained intact.

Task Description The Diapix task is a spot-the-difference task where two partners have similar but different pictures and must describe the pictures to each other to identify the differences (Figure 2.3). The pictures are themed either as Farm scene (shown below), Beach scene, or Street scene, with two stimulus sets selected from each theme. The partners cannot see each other's pictures and were seated in separate sound-isolated booths. They wore headsets with boom microphones and communicated through an open voice loop. Each set of pictures had 12 differences. Participants received instructions to find the 12 differences as quickly as possible through communication. In addition to verbal communication, one speaker was able to place visual markers on the map that would appear on both partners' maps. The intended purpose of the markers was to mark the locations of identified differences as a way to facilitate monitoring task status (i.e., counting how many differences were found), but the use of markers was not restricted. There was no limit to the number of markers that could be placed and they could be removed if placed mistakenly. Therefore, partners could use the markers as a temporary pointing device to allow for deixis, by adding and removing a marker (and indeed some partners used this strategy).

Measures The Diapix task used a completion time metric while fixing accuracy. Because every team did not find all the differences, task completion time was recorded after 8 of the 12 differences had been found. There was no penalty for locations that were incorrectly marked as differences.

Task Conditions This corpus was originally collected to examine how speech is adapted for complex acoustic environments, particularly when the interlocutors are in mis-matched acoustic environments. To manipulate this, the participants performed this task while being exposed to one of three possible background sound mixtures: quiet, sparse speech babble from two simultaneous talkers, or dense speech babble from eight simultaneous talkers. The 85 dB babble mixtures combined multiple recorded utterances from the coordinate response measure task (Bolia et al., 2000), which are structured radio-like phrases about a

color-number combination to a specific call sign (e.g., *'Ready Eagle go to Blue Six now'*).

During some conditions, partners received different background sound mixtures and occasionally received the same background sound mixture. These manipulations may have affected the dialogue interactions by increasing the likelihood of State 1 understanding (Table 1.3), where the addressee knows that the speaker provided some potentially misunderstood contribution. Eight pairs completed 1 trial of each of the 6 background sound mixture conditions for a total of 48 dialogues in the corpus. A transcript excerpt is shown in Table 2.5.

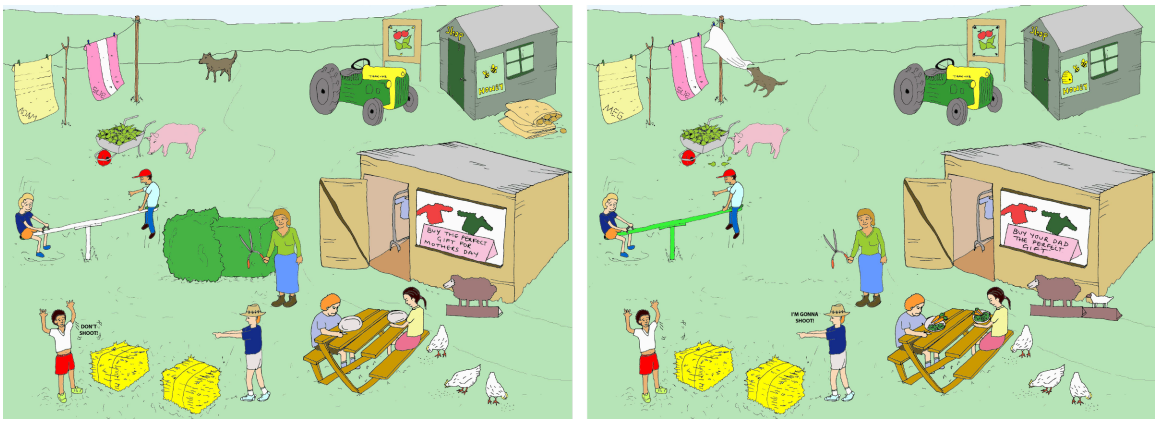

Figure 2.3: Two images from the Diapix task that illustrate the differences between the images that the partners were instructed to find through describing the pictures to each other. One partner would get the left image and one partner would receive the right image. For an example difference, the seesaw is colored white in the left image and colored green in the right image.

2.1.4 Combat Search and Rescue (CSAR) Task Corpus

The Combat Search and Rescue (CSAR) task is a team dialogue task collected by the Air Force Research Laboratory. The task was a team navigation task originally developed to examine the effect of spatialized radio displays on navigation and the role of landmarks in navigation dialogue (Hampton et al., 2012; Hampton, 2013). It was selected specifically for its task characteristics that were expected to demand articulation work.

Table 2.5: An excerpt from the AFRL Diapix task Corpus from a Farm scene. In this example, the background noise was at a high-level and some requests are made to repeat whole utterances (#9) or certain words (#15).

#	Speaker	Transcript
8	B	do you have two yellow hay patches or whatever hay bales
9	A	what'd you say
10	B	do you have two like yellow hay stacks
11	A	that's right yellow hay stacks that's what I got
12	B	okay
13	A	there's a uh lady pretending to shoot her with a blue shirt white shorts and like blue shoes
14	B	uh yeah I think my shoes are white
15	A	what color shoes you say
16	B	I think they're white
17	A	okay mine are like kinda bluish I don't know
18	B	okay mine might be kinda blue oh nope
19	A	alright there's two people sitting at the picnic table
20	B	does
21	A	oh one
22	B	hey hold on a sec does your girl say I'm gonna shoot
23	A	no
24	B	alright my girl says that

Participants There were 8 participants organized into 4 teams of 2 people. Teams remained intact throughout the corpus.

Task Description The team navigation task was conducted in a large urban virtual environment. Two participants were cooperating within the virtual environment while being located in separate but networked virtual reality facilities. On each trial, participants started in different locations in the virtual world (approximately opposite sides of the environment) and they had to rendezvous with each other as quickly as possible. Both participants 'moved' through the environment using a hand controller. Participants communicated with each other over a push-to-talk radio. Virtual terrain varied from trial to trial but was always measured 500 meters by 500 meters.

In addition to their navigation task, participants had to avoid getting shot. Enemy

forces were searching for both participants and could shoot them. This additional goal complicated the navigation task and provided perturbations to task progress and potential verbal exchange. It was expected that articulation work could result from the presence of enemy forces (e.g., discussion about a change in plan). Only one participant, the rescuer, had a rifle and could shoot back. The other participant, the 'to-be-rescued,' did not have a weapon and could only hide or flee. Figure 2.4 shows one virtual reality facility and illustrates an enemy on the right screen.

Participants were instructed to find each other as quickly as possible while avoiding getting shot.

Measures Completion time served as the performance metric. The trial ended when the two participants found each other, defined by coming within 3 meters of each other. Trials terminated if the team did not rendezvous within ten minutes. This was uncommon, 5 of the 120 trials failed to finish. (Team 3 failed to rendezvous 2 times; Team 4 failed to rendezvous 3 times). In these cases, ten minutes served as their completion time.

Task Conditions The corpus originally manipulated the type of radio display available to the team and the presence of landmarks in the virtual environment. The original corpus had a baseline monoaural radio communication condition (i.e., diotic presentation) and a spatialized radio condition. Prior research showed that the acoustic signal of the spatialized radio successfully conveyed spatial information, decreasing the navigation instructions and increasing the amount of task irrelevant dialogue (Hampton et al., 2012; Hampton, 2013). Because the primary interest of the current research was in task-related navigation dialogues, analyses only used the conditions with monoaural radio communications.

Landmarks On a given trial, landmarks were either present or absent. Landmarks were tall/visible, salient, distinct structures (e.g., mosque, water tower). Figure 2.5 shows an example landmark from the task.

Figure 2.4: A wide-angle picture from inside one of the virtual reality facilities used in this task. The screen on the right shows one of the enemy forces.

Figure 2.5: A screenshot from the CSAR task that illustrates a landmark, in this case, a conventional water tower.

Each of the 4 teams completed 15 trials with landmarks and 15 trials without landmarks. Of the 120 trials collected, 2 trials were missing completion time data due to a data logging error, leaving 118 trials for analysis. An excerpt is presented in Table 2.6.

Table 2.6: An excerpt from the CSAR task Corpus that illustrates how participants had to change plans due to the enemy forces.

#	Talker	Transcript
8	A	I'm heading to the far it looks like north east wall and I'm a head down the road
9	B	northeast wall and then
10	B	that's not you
11	A	yeah
12	B	let me know if you see that buddha statue or the park
13	A	yeah I see the edge of the map
14	A	I
15	B	over there this isn't very
16	A	see a buddha statue
17	B	big
18	B	alright
19	A	looks like the road I'm on is going to come out right in between that and the park
20	B	okay weird my road looks like that too
21	B	bunch of armed guys by this thing I don't know if we should meet
22	B	oh crap
23	A	gettin shot
24	B	oh yep he's an excellent shot too
25	A	no
26	B	he's hitting me every time I got a feeling you're right around this wall
27	B	God I need a gun yeah this this whole tower there's another guy with a gun
28	B	I need you to wipe these people out for me I'm gonna hide around this corner
29	B	God he got me again
30	A	I see you

2.2 Recurrence Analyses

Three models were tested in search of the model that best predicts task performance, following Fusaroli and Tylén (2016) as illustrated in Figure 1.2. The *alignment* model was represented by cross recurrence quantification analysis (CRQA) of a time series of Speaker A with a time series of Speaker B. The *coordination* model is represented by recurrence quantification analysis (RQA) of the time series for the entire block (Speaker A and Speaker B). A baseline *self-consistency* model is represented by performing RQA of each speaker's time series with his/herself and using the recurrence plot with the highest recurrence rate.

The analyses examined five levels: two lexical levels (word level and morpheme level), the syntactic level, the pitch level, and the rhythm level. The levels of analysis divide into categorical data analyses (i.e., morpheme, word, and syntactic) and continuous data analyses (i.e., pitch and rhythm). The recurrence metrics of Recurrence Rate (RR), Determinism (DET), Average Line Length (L) and Line Entropy (ENTR) were calculated for each level. Fusaroli and Tylén (2016) only used L and ENTR, but the current research added RR and DET because of prior use in recurrence analysis of language (e.g., Orsucci et al., 2013; Gorman et al., 2012; Coco et al., 2017). The current analysis allowed statistical significance to inform whether or not these predictors were relevant.

2.2.1 Categorical Data

The lexical levels used the orthographic transcriptions with partial words and punctuation removed. The word level analysis used single words as the unit of analysis and the morpheme level treated each character as one step in time, and later the embedding dimension described below combined characters into letter trigrams. The syntactic level resulted from the transcripts using part-of-speech (POS) tags output by the Stanford Natural Language Group's Log-linear Part-Of-Speech Tagger `http://nlp.stanford.edu/software/tagger.shtml`. The tagger output POS tags in the Penn Treebank format

Table 2.7: Example part-of-speech (POS) tags in the Penn Treebank format that are output by the Stanford Log-linear Part-Of-Speech Tagger.

Tag	Description
DT	Determiner
NN	Noun, common, singular or mass
NNP	Noun, proper, singular
RB	Adverb
VB	Verb, base form
VBD	Verb, past tense

(Taylor et al., 2003). Example tags appear in Table 2.7.

For the time series used in the categorical analyses, a single time step was a character, a word, or a POS tag depending on the level of analysis. There were no time steps counted for periods of silence where neither speaker was talking. Each item (character, word, or POS tag) was labeled with a unique numerical identifier. Overlapping speech was converted to sequential interleaved speech using the start time of each word. The alignment model time series were constructed by placing Speaker A's contributions into one time series and Speaker B's contributions into another time series. Two additional identifiers were created and added to the time series. One identifier was placed in Speaker A's time series to account for when Speaker A was silent while Speaker B was talking. The other identifier was placed in Speaker B's time series for when Speaker B was silent while Speaker A was talking. These silence identifiers were necessary to preserve the sequencing of the original dialogue, and thereby preserve the phase information of the time series. These silence identifiers were different so that no spurious recurrence between silences would be measured.

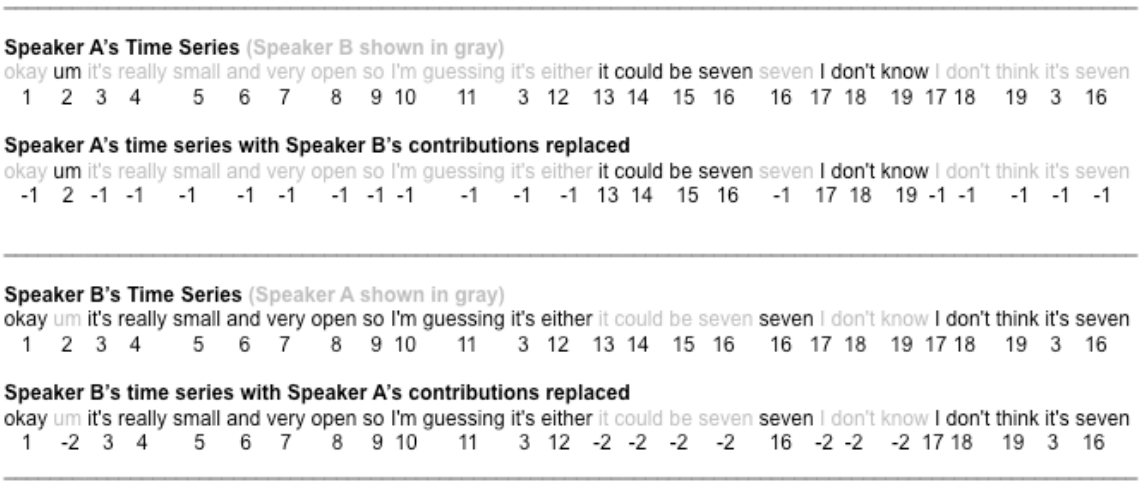

Figure 2.6: An illustration of how categorical time series were constructed for the alignment analysis, using an example from the word level of the Uncertainty Elicitation task. Speaker A is shown at top and Speaker B is shown at bottom. First, words were given a unique numerical identifier, shown here beneath each word. Then, the silence identifiers -1 and -2 were added to Speaker A's and Speaker B's respective time series. Note that there is no time step for which both Speaker A and Speaker B are silent.

2.2.2 Continuous Data

Regarding the continuous levels, the pitch level was generated using Praat (Boersma and Weenink, 2001), which extracts pitch information using an autocorrelation calculation that corrects for artifacts and octave jumps. The minimum pitch value was set to 70 Hz and the maximum value was set to 600 Hz. Pitch was originally sampled at a rate of 100 Hz, however some tasks required the data to be down sampled because calculating recurrence plots on long trials was consuming more RAM memory than was available (64 GB). Table 2.8 details the sample rates used in the different analyses. For constructing the alignment prosody time series, a multiple step process was used (Figure 2.7). First, the pitch information for each speaker was placed in separate time series. Then the silences were removed from each times series and the data were normalized to mean of 0 standard deviation of 1 for each speaker. Silences were then put back into the time series for periods of time when the other speaker was talking. Silence was coded as large negative values to avoid spurious recurrence with the normalized pitch data (-2000 or -3000 for silence in

Speaker A or B, respectively). The end result was normalized pitch information for each speaker. The sequencing of pitch information was maintained and all periods where both speakers were silent removed.

Coordination prosody time series were constructed in a similar fashion. The silence was removed and both time series were normalized, then the two time series were interleaved maintaining the original sequencing. When there was overlapping speech between the speakers, the higher post-normalization value was chosen. For the baseline prosody analysis, silences were removed from each speakers' time series.

The rhythm level was generated by discretization of the pitch trace into silent intervals and speech intervals. Silence was defined as the absence of pitch for 20 ms (2 samples).

2.2.3 Recurrence Parameters

Prior to calculating the recurrence plot to derive the recurrence metrics, RQA and CRQA require a number of parameters. Parameter values were set and recurrence plots were calculated keeping with Fusaroli and Tylén (2016), clarified and confirmed (R. Fusaroli, personal communication, Aug. 5, 2017). These varied between the categorical and continuous data analyses. For the categorical data, the radius value was set to 0. This meant for nominal data only an exact match was counted as a recurrence.

The threshold for a line (parallel to the positive diagonal) was set at 2. Time delay was set to 1. The word level analysis and syntactic level analysis used an embed value of 1. The morpheme level analysis used an embed value of 3. The continuous data analysis differed from the categorical data analysis. The delay value was set by calculating the mutual information for each model's time series, finding the first local minimum in each, then taking the largest value of the three models. For the pitch level and rhythm level analyses, the embedding dimension was estimated using a false nearest neighbor method for each model, again taking the largest value of the three models. Table 2.8 summarizes

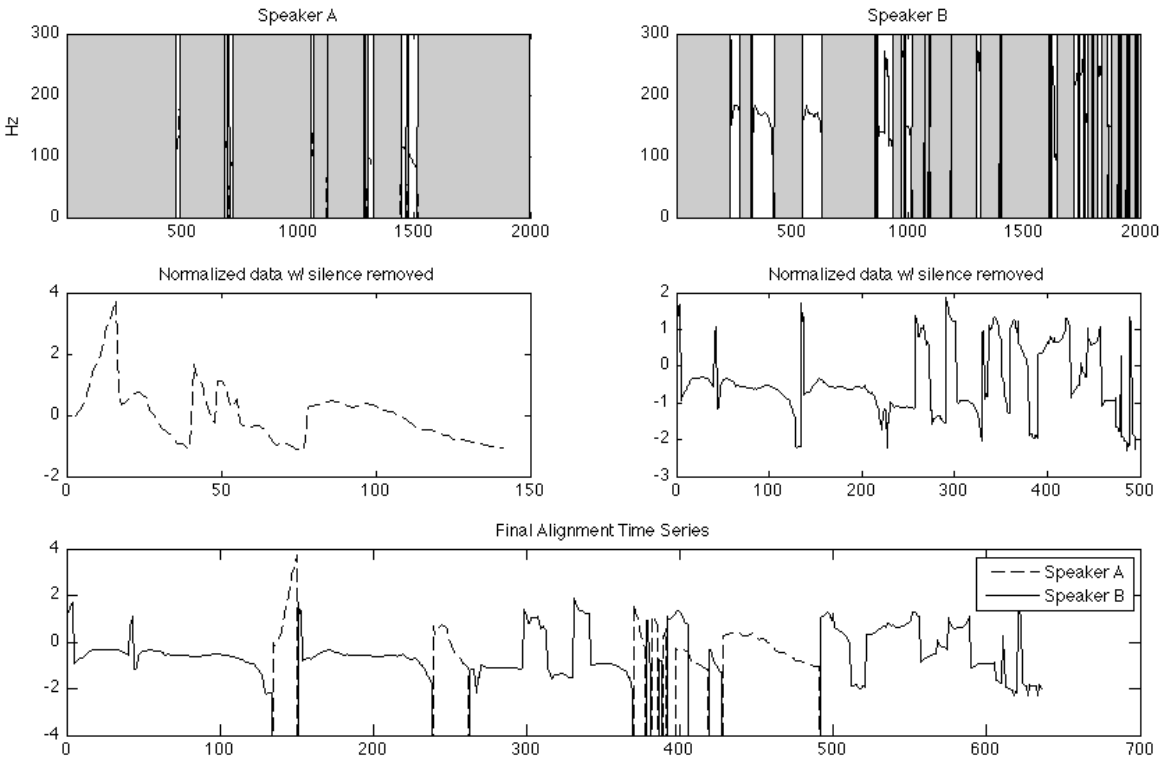

Figure 2.7: An illustration of how prosodic time series were constructed for the alignment analysis. The X axis in each panel shows the sample position. At top, a 2000-sample pitch trace for each speaker. Silences were identified (shown in gray) and removed from each speaker. At middle, the data were much shorter and were normalized. At bottom, both final time series are shown overlaid on the same axes. Some silence was put back with placeholder values that preserved sequencing. Periods where both speakers were silent were eliminated, as indicated by the shortened time series.

Table 2.8: Recurrence quantification analysis values used in the continuous data analyses. Smaller sample rates had to be used due to practical limitations.

Task	Level	Sample Rate	Delay	Embed
Map Task	Prosody	33	9	4
	Rhythm	25	2	4
Uncertainty Task	Prosody	100	2	4
	Rhythm	25	2	4
Diapix Task	Prosody	100	2	9
	Rhythm	50	2	4
CSAR Task	Prosody	100	9	4
	Rhythm	50	2	4

the delay and embedding values used.

2.2.4 Track 2 Dialogue

To facilitate analysis of the coordination model's contents, a model representing Track 2 dialogue was developed using the Linguistic Inquiry and Word Count 2015 (LIWC) text analysis program (Pennebaker et al., 2015). The analysis relied on two word lists that may capture Track 2 issues of dialogue management: $Assent$ (e.g., agree, OK, yes) and $Certainty$ (e.g., must, specific, clear). Analysis focused on these two word lists because they may capture acknowledging understanding (e.g., backchannel communication) and signaling non-understanding (e.g., clarification requests). LIWC analyzed the presence of *Assent* and *Certainty* by counting instances of *Assent* and *Certainty* lexicons in each trial of each task. The counts were normalized by dividing each count by the word count for that trial.

2.2.5 Description of Analyses

Individual recurrence metrics of recurrence rate (RR), determinism (DET), average line length (L), and entropy (ENTR) were calculated from recurrence plots created for each of the three models (i.e., alignment, coordination, baseline), for each level of analysis, and for each trial in each corpus. The following analyses used these recurrence metrics to address a number of questions. Prior to analysis, recurrence metrics were tested for measuring structure over and above what recurrence would be expected by chance (Chapter 3).

The first set of analyses investigated the ability of the alignment, coordination, and baseline models to predict performance following Fusaroli and Tylén (2016) (Chapter 4). Subsequent analyses accounted for the repeated measures nature of the data. The second set of analyses sought to examine the contents of the coordination recurrence model through statistical mediation (Chapter 5). Mediation analyses tested if the LIWC model's relationship to performance was mediated by the coordination model, thereby indicating that coordination captures Track 2 dialogue and is the strategic variant. The third set of analyses investigated contents of the recurrence models through measuring the relationships between the models (Chapter 6). As alignment and coordination are alternative accounts of grounding phenomenon, it was expected that models representing each theory should not be related to each other. The fourth set of analyses further examined the contents of the recurrence models using the accuracy and time performance measures (Chapter 7). Here, an additional accuracy metric was extracted from the Uncertainty task to test prediction of accuracy more generally. In addition, a model of grounding should represent a time series's content rather than its length, so analyses examined the relationship between recurrence metrics and word count. The fifth and final set of analyses tested the relationship between the recurrence metrics and previously introduced methods of characterizing communication including articulation work (Chapter 8).

Analyses of the continuous data were performed using the Matlab CRP Toolbox (Marwan et al., 2007), available at `http://tocsy.pik-potsdam.de/CRPtoolbox/`.

Analyses of the categorical data were performed in R, using the $crqa$ package described in Coco and Dale (2014). All other statistical analyses were performed in R. A portion of the tables were created with the help of the $stargazer$ R package (Hlavac, 2018).

Results: Chance Analysis

Prior to the analyses, it is good practice to test that the structure of recurrence represented by these metrics is not due to chance (Fusaroli and Tylén, 2016; Louwerse et al., 2012). This assures that the values are not the spurious result of repetition inherent in the English language or the task domain. The null hypothesis is that values are randomly drawn from a uniform distribution, which can be approximated by randomly shuffling the time series (Kantz and Schreiber, 2004). Two-sided paired *t*-tests compared the recurrence metrics from 'forwards' time series to the recurrence metrics from the randomly shuffled time series for each model for each level of analysis. For the categorical levels of analysis (morpheme, word, and syntax), the words in the time series were randomly ordered for each of the three models. For the continuous levels of analysis (prosody and rhythm), the samples in the time series were randomly ordered for each of the three models. Using shuffled time series is a conservative procedure compared to other alternatives because a shuffled time series preserves the recurrence rate better than alternative control methods (Louwerse et al., 2012), as was shown in the Green Eggs demonstration above (Table 1.2). When recurrence rate is higher, diagonal line structures are more likely to occur by chance than other control methods. In all, 240 tests were conducted (4 tasks x 3 models x 4 recurrence metrics x 5 levels of analysis).

For the majority of tasks and models, results indicated that recurrence structure was significantly different from shuffled controls (see Appendix A). Some metrics were not different from chance, as anticipated. Expected consistency with chance was found for

the pitch-level, word-level and syntax-level recurrence rate (RR) metrics. For the prosody-level analyses, recurrence rate was intentionally set to 4% by varying the radius parameter, leading to little variation between the analyses and shuffled controls. For the word-level and syntax-level analyses, recurrence was calculated matching individual words (embed value of 1), therefore reordering but not adding to or removing from the time series results in recurrence rates that are exactly the same for forwards and shuffled time series. (The morpheme-level control is different because those metrics result from letter trigrams—an embed value of 3).

However, unanticipated consistency with chance occurred for metrics in the morpheme-level analyses for the Diapix and CSAR tasks, precluding further analyses of recurrence for those tasks (Table 3.1). The Diapix task Alignment model Determinism (DET) metric and the Baseline model Entropy (ENTR) metric were not significantly different from chance ($t(47) = -1.59$, $p = 0.12$; & $t(47) = 1.09$, $p = 0.28$, respectively). The CSAR task Alignment model had three metrics that were not significantly different from chance: DET $t(119) = -1.13$, $p = 0.18$, L $t(119) = 0.93$, $p = 0.35$, and ENTR $t(119) = 1.13$, $p = 0.26$. Further analyses included the metrics that were expected to be consistent with chance (word-level, syntax-level, and prosody-level RR), but excluded the unexpected metrics that were not significantly different from chance.

Table 3.1: Table showing p-values of chance analyses where recurrence measured was not significantly different from shuffled time series (shown in bold).

	Diapix Task - Morpheme Level			
	RR	**DET**	**L**	**ENTR**
Alignment	$< .001$	**0.12**	$< .001$	$< .001$
Coordination	$< .001$	$< .001$	$< .001$	$< .001$
Baseline	$< .001$	$< .001$	$< .001$	**0.28**
	CSAR Task - Morpheme Level			
Alignment	$< .001$	**0.18**	**0.35**	**0.26**
Coordination	$< .001$	$< .001$	$< .001$	$< .001$
Baseline	$< .001$	$< .001$	$< .001$	$< .001$

Results: Predicting Task Performance

These analyses investigated if the alignment, coordination, and baseline models predicted task performance. The three separate recurrence models output separate recurrence metrics used in different regression models, in order to assess the relationship of each recurrence model to task performance. This analysis process differs from a typical regression procedure where predictors are added or removed from a single regression model. Here, three regression models used predictors from different recurrence calculations. Recurrence metrics served as predictors for a linear regression model on the performance scores (i.e., completion times or accuracy) for each model for each corpus. Models were evaluated through examining *Adjusted* R^2 (*Adj* R^2) values, which is the proportion of variance explained but adjusted for the number of predictors. This facilitated comparisons between models with different numbers of predictors, as well as comparisons between the current research and Fusaroli and Tylén (2016). These models followed the analysis procedures from Fusaroli and Tylén assuming data from a between-subjects design. Subsequent tests addressed the repeated measures nature of this data, testing for both team effects and learning effects (Sections 4.2 and 4.3).

In sum, the coordination model had stronger relationships to task performance than the alignment model for most tasks and levels of analysis. After learning effects were statistically accounted for, the coordination model accounted for variance in performance whereas the alignment model did not. Moreover, after the differences between the teams were statistically accounted for, the coordination model accounted for variance in perfor-

mance whereas the alignment model did not.

4.1 Predicting Task Performance

The recurrence metrics served as predictors of performance in linear regression models for each task, for each model, and for each level of analysis. Outliers in the dependent measures were excluded casewise. Outliers were defined as any value below the first quartile (Q1) or above the third quartile (Q3) by 1.5 times the inner-quartile range (Q1-Q3). The distributions of dependent measures appear in Figure 4.1. Outliers appear as open circles.

A summary of variance in performance explained appears for each task in Figures 4.2, 4.3, 4.4, 4.5, and 4.6. Appendix B shows all model details. Across levels of analysis and tasks, more of the coordination models were significant predictors of performance than the alignment models (23 of 25 and 9 of 25 models, respectively). A two-sample test of the proportions showed that coordination models were significant more often than alignment models, $\chi^2(1) = 14.67$, $p < .001$. In addition, the coordination models explained more variance in performance than the alignment models. The baseline models were also significant more frequently than alignment models (19 of 25 and 9 of 25 models, respectively). A two-sample test of the proportions also showed that baseline models were significant more often than alignment models, $\chi^2(1) = 6.57$, $p < .05$. Moreover, baseline models predicted performance better than the alignment models. The baseline models accounted for a similar amount of performance as the coordination models in some cases, such as in the Uncertainty Elicitation task (Figure 4.2).

For each task, the single model that predicted the most performance variance (independent of level of analysis) was always a coordination model. This was a morpheme-level

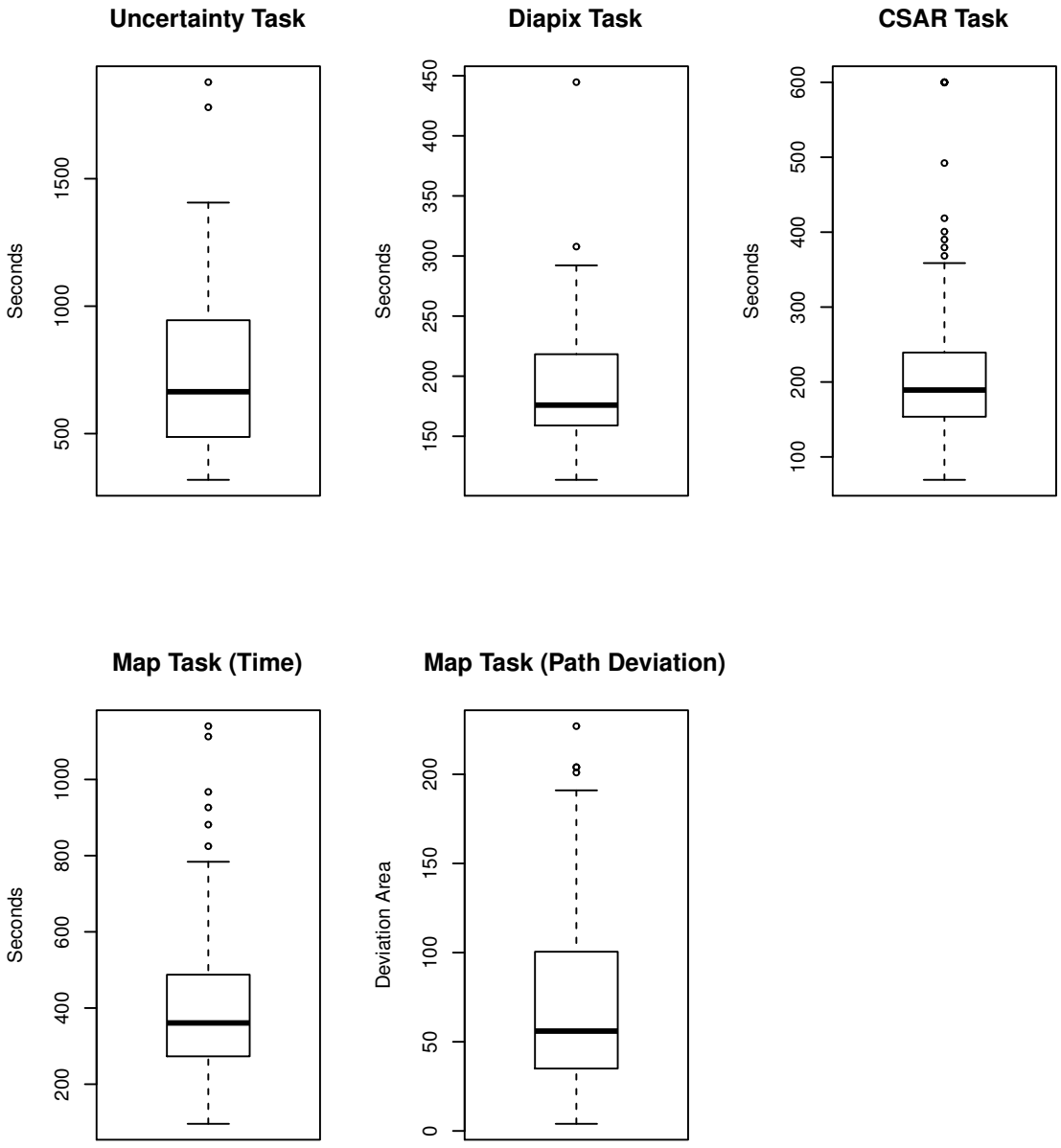

Figure 4.1: Box plots of dependent measures for each task. Seconds are shown for all plots except Path Deviation. The box indicates the inter-quartile range (Q1-Q3) and the bold line indicates the median value. The whiskers indicate 1.5 times the inter-quartile range and outliers appear as open circles.

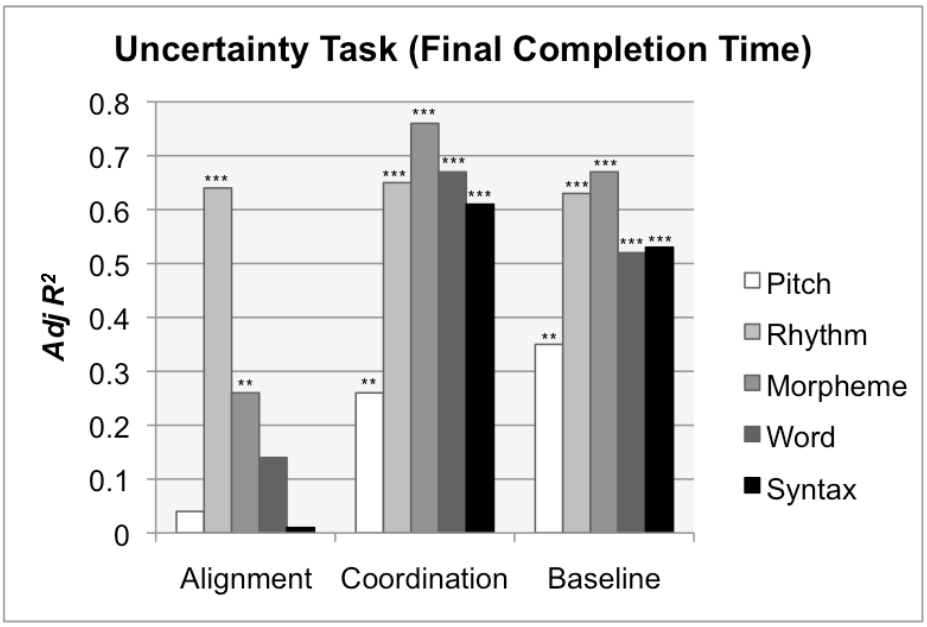

Figure 4.2: Overview of recurrence models prediction of final task completion times in the Uncertainty Elicitation task. (* $p < .05$; ** $p < .01$; *** $p < .001$)

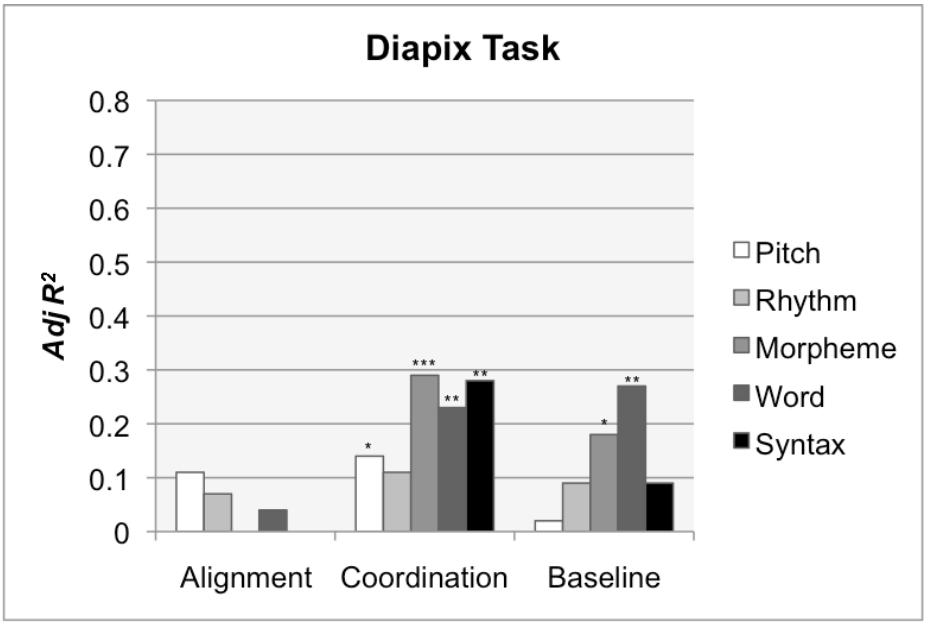

Figure 4.3: Overview of recurrence models' prediction of final task completion times in the Diapix task. (* $p < .05$; ** $p < .01$; *** $p < .001$)

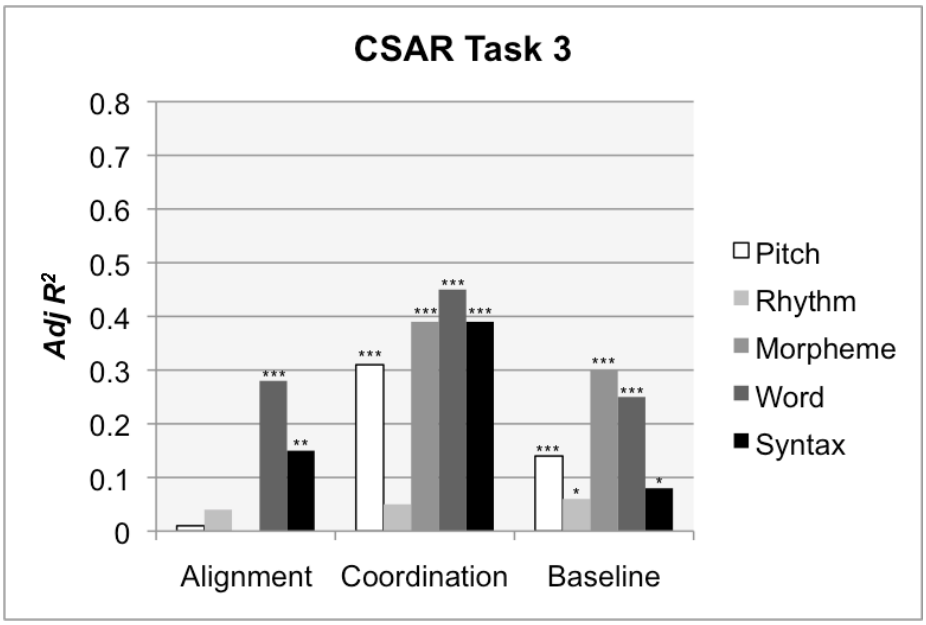

Figure 4.4: Overview of recurrence models' prediction of task completion times in the CSAR task. (* $p < .05$; ** $p < .01$; *** $p < .001$)

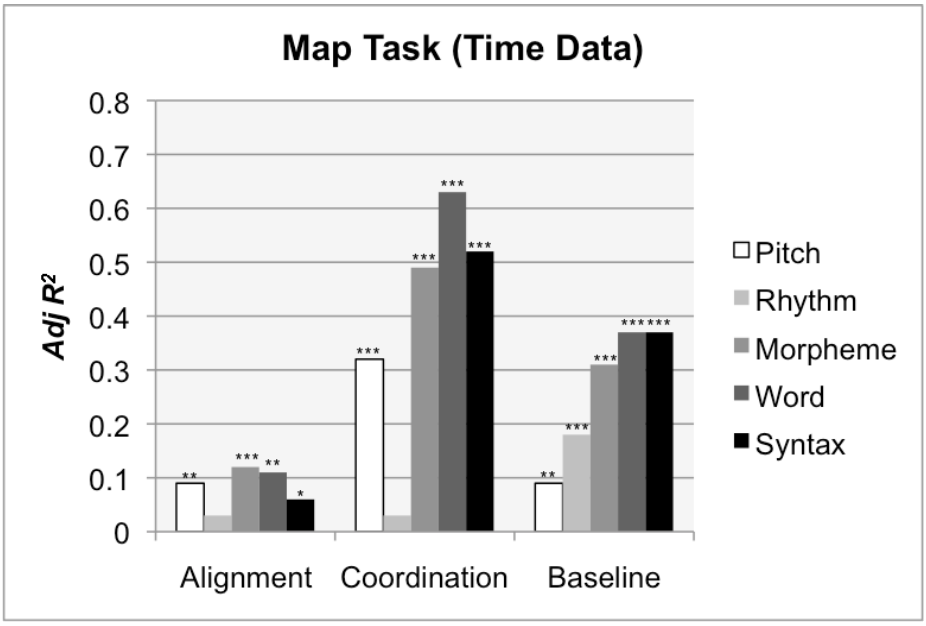

Figure 4.5: Overview of recurrence models prediction of time to completion in the Map task. (* $p < .05$; ** $p < .01$; *** $p < .001$)

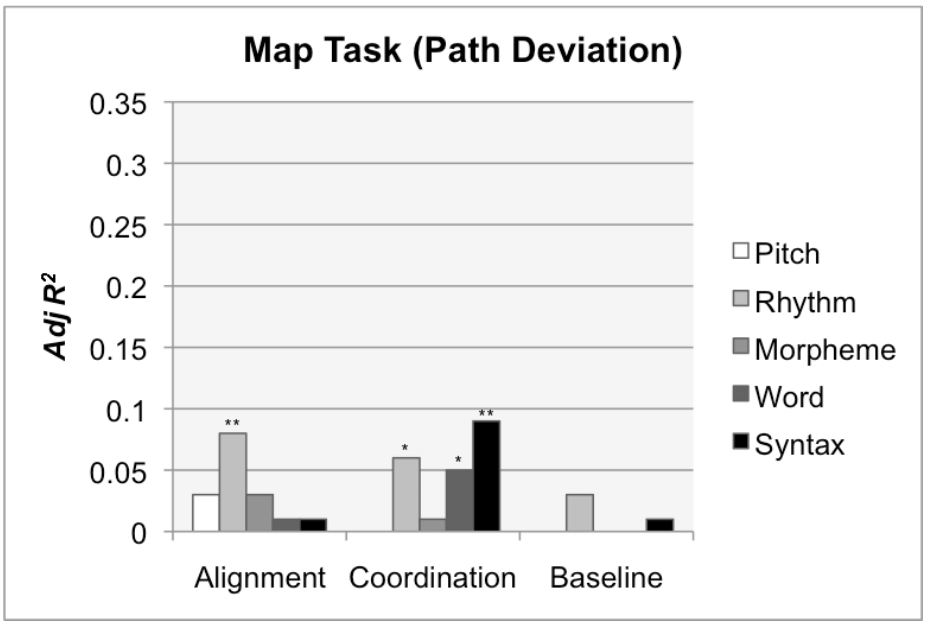

Figure 4.6: Overview of recurrence models' prediction of path deviation score in the Map task. Note: the ordinate range differs from the other plots. (* $p < .05$; ** $p < .01$)

model for the Uncertainty and Diapix tasks, a word-level model for the CSAR and Map task (time) tasks, and a syntax-level model for the Map task (path deviation). However, different amounts of variance were accounted for between the different tasks. For instance, the largest *Adjusted* R^2 value for the Uncertainty task was 0.76, for the CSAR task it was 0.45, whereas for the Diapix task it was 0.29. Notably, all models for the Map task (where accuracy is readily available) had lower *Adjusted* R^2 values when accuracy was the performance measure rather completion time. This is clearly shown by contrasting the pattern of findings in the Map task (Figure 4.5 and Figure 4.6; note the different ordinate scales). The best model predicting accuracy for the Map task was the syntax level coordination model (*Adj* $R^2 = 0.09$) compared to the time data's word level coordination model (*Adj* $R^2 = 0.63$). (A portion of these analyses were conducted on another accuracy measure extracted from the Uncertainty task, Section 7.1).

4.2 Learning Effects

Additional analyses addressed the repeated-measures nature of these corpora by testing for learning effects. Overall, models that were previously significant were significant after learning was statistically controlled with the exception of the CSAR task where most models lost significance. Details of statistical tests appear in Appendix C.

4.2.1 Uncertainty Elicitation Task

For the Uncertainty Elicitation task, learning analyses tested for improvement in (final) completion times in the 8 trials that each team performed (n = 40). A one-way analysis of variance (ANOVA) with block number as the only factor showed that completion times did not significantly change, $F(7, 31) = 0.61$, $p = 0.74$. Post hoc correlations between completion time and block number were conducted for each team (Table 4.1). Only Team 5 showed significant learning (r = -0.71, $t(6)$ = -2.45, $p < .05$), however other teams had moderate correlations that might not have been significant due to small power. Team 2 (r = -0.55) and Team 4 (r = +0.58) had correlations of similar magnitude but in opposite directions.

Table 4.1: Correlations testing for learning in the Uncertainty Elicitation task. All df = 6.

Team #	***r***	***t*-value**	***p*-value**
1	+0.12	0.29	0.78
2	-0.55	-1.59	0.16
3	-0.12	-0.29	0.78
4	+0.58	1.76	0.13
5	-0.71	-2.45	$< .05$

As a precaution, recurrence models were re-tested using the residual completion time after regressing block number on completion time.[1] In summary, all models that were

[1]The 7 degrees of freedom from block number was removed from subsequent error degrees of freedom.

prior significant predictors of performance remained significant and accounted for similar amounts of variance. The surprising exception was that the rhythm-level alignment model's predictions *increased*, *Adj* R^2 = 0.78, $F(4, 28) = 28.80$, $p < .001$. No models gained significance that originally failed to reach significance.

4.2.2 Diapix Task

For the Diapix task, learning analyses tested for improvement in completion times over the 6 trials that each team completed (n = 48). A one-way ANOVA with block number as the only factor showed that completion times did not significantly change, $F(5, 42) = 1.88$, $p = 0.12$, $\eta_p^2 = 0.18$. Post hoc correlations between completion time and block number were conducted for each team (Table 4.2). Only Team 4 showed significant learning (r = -0.94, $t(4)$ = -5.33, $p < .01$), however many other teams showed negative correlations and power was limited by the small number of blocks.

Table 4.2: Correlations testing for learning in the Diapix task. All df = 4.

Team #	***r***	***t*-value**	***p*-value**
1	-0.75	-2.23	0.09
2	-0.49	-1.12	0.33
3	+0.46	1.05	0.35
4	-0.94	-5.33	$< .01$
5	+0.06	0.13	0.90
6	-0.69	-1.91	0.13
7	-0.21	-0.43	0.69
8	-0.49	-1.11	0.33

As a precaution, recurrence models were re-tested using the residual completion time after regressing block number on completion time. In summary, models that were prior significant predictors of performance remained significant with the exception of the pitch-

Other tasks' degrees of freedom were similarly reduced by the appropriate amount.

level coordination model ($F(4, 37) = 0.85$, $p = 0.50$). Many models remained significant and increased in *Adjusted* R^2. The morpheme-level, word-level, and syntax-level coordination models remained significant, *Adj* $R^2 = 0.36$, $F(4, 37) = 6.75$, $p < .01$; *Adj* $R^2 = 0.25$, $F(4, 37) = 4.32$, $p < .01$; and *Adj* $R^2 = 0.25$, $F(4, 37) = 4.87$, $p < .01$, respectively. The morpheme-level and word-level baseline models remained significant, *Adj* $R^2 = 0.16$, $F(4, 37) = 2.87$, $p < .05$; and *Adj* $R^2 = 0.21$, $F(4, 37) = 3.68$, $p < .05$, respectively. One model became significant, the morpheme-level alignment model, *Adj* $R^2 = 0.18$, $F(4, 37) = 3.16$, $p < .05$.

4.2.3 CSAR Task

For the CSAR task, learning analyses tested for improvement in rendezvous times in the 30 trials that each team performed (n = 117). A one-way ANOVA with block number as the only factor showed rendezvous time did not significantly change, $F(29, 88) = 1.22$, $p = 0.24$. Post hoc correlations between completion time and block number were conducted for each team (Table 4.3). Team 2 was marginally significant, $r = -0.37$, $t(26) = -1.99$, $p = 0.057$.

As a precaution, recurrence models were re-tested using the residual completion time after regressing block number on completion time. All models that were previously significant lost significance: word-level and syntax-level alignment models, pitch-level, morpheme-level, word-level, and syntax-level coordination models, as well as the pitch-level, rhythm-level, morpheme-level, word-level, and syntax-level baseline models. One model gained significance after controlling for learning, the pitch-level alignment model, *Adj* $R^2 = 0.09$, $F(4, 59) = 2.96$, $p < .05$.

Table 4.3: Correlations testing for learning in the CSAR task. Degrees of freedom (df) varied between teams due to missing data.

Team #	*r*	*t*-value	df	*p*-value
1	+0.34	1.69	22	0.10
2	-0.37	-1.99	26	0.06
3	+0.04	-0.29	26	0.86
4	-0.16	-0.79	24	0.44

4.2.4 Map Task: Completion Time

For the Map task, each team only completed two trials but teams were intermixed in the experimental design to manipulate familiarity. As a result, each participant completed 4 trials, 2 in a familiar team and 2 in an unfamiliar team. Learning analyses tested for improvements in completion time scores in the 4 trials that each participant completed (n = 128). Six outliers were removed, resulting in 122 trials for analysis. A one-way ANOVA with block number as the only factor showed that completion time did not change significantly, $F(3, 118) = 2.58$, $p = 0.056$, $\eta_p^2 = 0.06$. Post-hoc correlations for each team are omitted because teams were intermingled throughout the experiment and each pair of participants only completed 2 trials together.

As a precaution, recurrence models were re-tested using the residual completion time after regressing block number on completion time. All the models remained significant after controlling for possible learning. The alignment models, the coordination models and the baseline maintained the amount of variance explained, changing less than 3%.

4.2.5 Map Task: Path Deviation

Learning analyses tested for improvements in path deviation scores in the 4 trials that each participant completed (n = 128). Four outliers were removed, resulting in 124 trials for analysis. A one-way ANOVA with block number as the only factor showed that path devi-

ation did change significantly, $F(3, 120) = 6.60$, $p < .001$, $\eta_p^2 = 0.14$. Post-hoc correlations for each team are omitted because teams were intermingled throughout the experiment and each pair of participants only completed 2 trials together.

To account for learning, Map task models were re-tested using the residual path deviation after regressing block number on path deviation. In summary, the models that were prior significant predictors of performance remained significant with one exception. In addition, four models became significant after controlling for learning. The rhythm-level coordination model lost significance, $Adj\ R^2 = 0.05$, $F(4, 116) = 2.43$, $p = 0.051$. The rhythm-level alignment model remained a significant predictor of residual path deviation, $Adj\ R^2 = 0.07$, $F(4, 116) = 3.37$, $p < .05$. The word-level and syntax-level coordination models remained significant predictors of residual path deviation as well, $Adj\ R^2 = 0.07$, $F(4, 116) = 3.19$, $p < .05$, and $Adj\ R^2 = 0.13$, $F(4, 116) = 5.26$, $p < .001$, respectively. The morpheme-level coordination model became significant after controlling for learning, $Adj\ R^2 = 0.05$, $F(4, 116) = 2.54$, $p < .05$. Three baseline models became significant after controlling for learning, the rhythm-level baseline model ($Adj\ R^2 = 0.05$, $F(4, 116) = 2.75$, $p < .05$), the word-level baseline model ($Adj\ R^2 = 0.05$, $F(4, 116) = 2.67$, $p < .05$), and the syntax-level baseline model ($Adj\ R^2 = 0.06$, $F(4, 116) = 2.72$, $p < .05$).

4.3 Team Performance

Due to the repeated-measures nature of these corpora, additional analyses tested for the team contribution to performance when possible. Each analysis was an ANOVA with team ID as the only factor.

4.3.1 Map Task

In the Map task, team differences were not tested as teams were blended in the experiment and each team only completed 2 trials together.

4.3.2 Diapix Task

The team differences in the Diapix task were not significant, $F(7, 40) = 0.95$, $p = 0.48$.

4.3.3 CSAR Task

The team differences in the CSAR task were not significant, $F(3, 102) = 2.45$, $p = .06$, $\eta_p^2 = 0.07$.

4.3.4 Uncertainty Task

In the Uncertainty task, the team differences were significant, $F(4, 35) = 6.04$, $p < .001$, $\eta_p^2 = 0.41$. As Figure 4.7 indicates, Team 2 and Team 3 contributed to the large effect size. Using model comparison, we tested if each recurrence model could explain variance over and above the team ID factor.

For the rhythm-level, the alignment model remained significant, $\Delta R^2 = 0.45$, $F(4, 31) = 20.19$, $p < .001$. For the morpheme-level, the alignment was no longer significant, $\Delta R^2 = 0.05$, $F(4, 31) = 1.75$, $p = .16$.

For the pitch-level and rhythm-level, the coordination model remained significant, $\Delta R^2 = 0.13$, $F(4, 29) = 2.91$, $p < .05$, and $\Delta R^2 = 0.30$, $F(4, 31) = 8.21$, $p < .001$, respectively. For the morpheme-level, word-level and syntax-level, the coordination model remained significant, $\Delta R^2 = 0.30$ $F(4, 31) = 8.21$, $p < .001$; $\Delta R^2 = 0.34$, $F(4, 31) = 13.72$, $p < .001$; and $\Delta R^2 = 0.26$ $F(4, 31) = 6.78$, $p < .001$, respectively.

The baseline models also remained significant at all levels of analysis: pitch-level

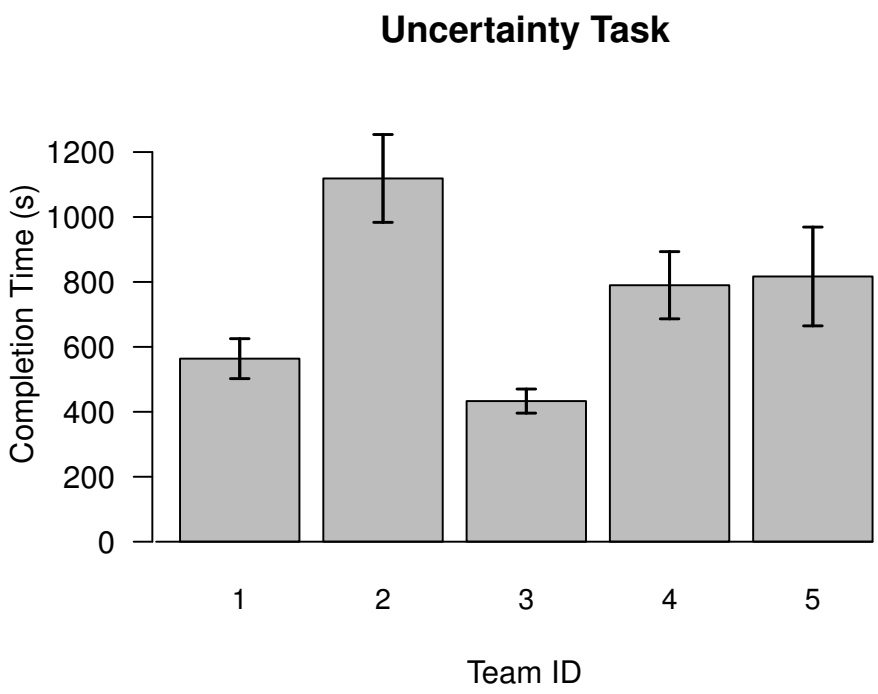

Figure 4.7: Mean team completion times for the Uncertainty Task. Error bars show $\pm$ 1 standard error.

(ΔR^2 = 0.11 F(4, 31) = 2.80, $p < .05$), rhythm-level (ΔR^2 = 0.39 F(4, 31) = 13.59, $p <$ .001), morpheme-level (ΔR^2 = 0.32 F(4, 31) = 9.21, $p < .001$), word-level (ΔR^2 = 0.17 F(4, 31) = 4.10, $p < .001$), and syntax-level (ΔR^2 = 0.20 F(4, 31) = 5.04, $p < .01$).

Results: Mediation Analyses

Statistical mediation aided in interpreting the results of the coordination model. These analyses examined the construct validity of the coordination recurrence model by testing for a relationship with Track 2 dialogue (Clark, 1996). Track 2 dialogue was estimated with the Linguistic Inquiry and Word Count 2015 (LIWC) text analysis program (Pennebaker et al., 2015), which uses pre-defined word lists that measure different dimensions of text. The analysis relied on two pre-defined word lists that may capture Track 2 issues of dialogue management: $Assent$ (e.g., agree, OK, yes) and $Certainty$ (e.g., must, specific, clear). Analysis focused on these two word lists because they may capture acknowledging understanding (e.g., backchannel communication) and signaling non-understanding (e.g., clarification requests). LIWC analyzed the presence of *Assent* and *Certainty* by counting instances of *Assent* and *Certainty* lexicons in each trial of each task. LIWC normalizes the counts by dividing each count by the word count for that trial.

Mediation following Baron and Kenny (1986) involved three "Steps" where the LIWC categories were treated as independent variables (IVs) and the recurrence parameters were treated as mediators (Ms): Step 1) the IVs and performance, Step 2) the IVs and the Ms, and Step 3) the (IVs + Ms) and performance. Complete mediation occurs when the IVs are related to the Ms and related to performance in the absence of the Ms, yet unrelated to performance when Ms are present. Partial mediation occurs when there is still a significant relationship between IVs and performance in the presence of Ms, but the relationship is reduced. Multiple linear regression was used for Steps 1 and 3 while multivariate analysis

of variance (MANOVA) was used for Step 2 in order to test for a relationship between multiple LIWC categories and multiple recurrence-parameter mediators. The LIWC model is based on word counts, so the word-level coordination recurrence model was used for all mediation analyses.

In summary, coordination completely mediated the LIWC model's relationship to task completion time in the Uncertainty task and partially mediated it's relationship in the Map task. However there was no mediation for the Diapix task, the CSAR task or the Map task path deviation measure.

5.1 Uncertainty Elicitation Task

For the Uncertainty Elicitation task, the coordination model completely mediated LIWC's relationship to task completion times (Table 5.1). At Step 1, the LIWC model was significantly related to performance, *Adj* R^2 = 0.43, $F(2, 37) = 15.42$, $p < .001$. Both *Assent* and *Certain* were significant predictors in the model, β = -.76, $t(37)$ = -5.48, $p < .001$; and β = -.49, $t(37)$ = -3.48, $p < .01$, respectively. At Step 2, both *Assent* and *Certain* were significantly related to the coordination model, $F(1, 4) = 6.35$, $p < .001$; and $F(1, 4) = 9.74$, $p < .001$. At Step 3, both *Assent* and *Certain* ceased to be significant in the presence of the coordination model, β = -.18, $t(33)$ = -1.06, p = .29; and β = .03, $t(33)$ = 0.16, p = .87.

Table 5.1: Uncertainty task mediation analysis for LIWC—See text for details. (*$p < .05$, **$p < .01$, ***$p < .001$)

Step 1— LIWC's relation to performance			
Assent***	$< .001$	Certain**	$< .01$
Step 2— LIWC's relation to Coordination			
Assent***	$< .001$	Certain***	$< .001$
Step 3—LIWC's & Coordination's relation to performance			
Assent	0.30	Certainty	0.89
RR_C*	$< .05$	L_C**	$< .01$
DET_C	0.30	$ENTR_C$	0.32

5.2 Map Task

For the Map task's completion time measure, the coordination model partially mediated LIWC's relationship to task completion times (Table 5.2). The LIWC model was a significant predictor of performance, *Adj* R^2 = 0.06, $F(2, 119) = 4.92$, $p < .01$. The *Certain* list was related to performance (β = +.26, $t(119) = 2.90$, $p < .01$). However the *Assent* list was not related to performance (β = -.12, $t(119)$ = -1.35, $p = 0.18$). At Step 2, only the *Certain* list was related to the coordination model, $F(4, 122) = 3.38$, $p < .05$. In the presence of the coordination model, *Certainty* was no longer related to performance, $\beta = 0.0$, $t(115) = 0.02$, $p = 0.98$.

For the Map task's path deviation measure, the LIWC model was not related to path deviation scores (*Adj* R^2 = 0.00, $F(2, 125) = 0.08$, $p = .93$) and therefore the coordination model did not mediate LIWC's relationship to performance (Table 5.3). The *Certain* list was related to the coordination model ($F(4, 122) = 3.38$, $p < .05$), but the *Assent* list was not related ($F(4, 122) = 1.01$, $p = 0.41$).

Table 5.2: Map task completion time mediation analysis for LIWC—See text for details. (*$p < .05$, **$p < .01$, ***$p < .001$)

Step 1— LIWC's relation to performance			
Assent	0.18	Certain**	$< .01$
Step 2— LIWC's relation to Coordination			
Assent	0.41	Certain*	$< .05$
Step 3—LIWC's & Coordination's relation to performance			
Assent	0.20	Certainty	0.98
RR_C***	$< .001$	L_C***	$< .001$
DET_C**	$< .01$	$ENTR_C$***	$< .001$

Table 5.3: Map task path deviation mediation analysis for LIWC. Step 3 was not conducted because Step 1 was not successful—See text for details. (*$p < .05$)

Step 1— LIWC's relation to performance			
Assent	0.96	Certain	0.70
Step 2— LIWC's relation to Coordination			
Assent	0.41	Certain*	$< .05$

5.3 Diapix Task

For the Diapix task, the analyses did not accomplish Step 1. The LIWC model was not related to completion times (*Adj* R^2 = 0.00, $F(2, 45) = 0.99$, $p = .38$) and therefore the coordination model did not mediate LIWC's relationship to performance (Table 5.4). *Assent* was related to the coordination model ($F(4, 42) = 3.11$, $p < .05$), but *Certainty* was not related ($F(4, 42) = 1.79$, $p = .15$).

Table 5.4: Diapix task mediation analysis for LIWC. Step 3 was not conducted because Step 1 was not successful—See text for details. (*$p < .05$)

Step 1— LIWC's relation to performance			
Assent	0.18	Certain	0.57
Step 2— LIWC's relation to Coordination			
Assent*	$< .05$	Certain	0.15

5.4 CSAR Task

For the CSAR task, the analyses did not accomplish Step 1. The LIWC model was not related to completion times ($Adj\ R^2 = 0.00$, $F(2, 107) = 1.03$, $p = .36$) and therefore the coordination model did not mediate LIWC's relationship to performance (Table 5.5). Neither the *Assent* or *Certain* lists were related to the coordination model, $F(4, 114) = 0.01$, $p = 0.93$; and $F(4, 114) = 0.01$, $p = 0.83$.

Table 5.5: CSAR task mediation analysis for LIWC. Step 3 was not conducted because Step 1 was not successful—See text for details.

Step 1— LIWC's relation to performance			
Assent	0.12	Certain	0.67
Step 2— LIWC's relation to Coordination			
Assent	0.93	Certain	0.83

5.5 LIWC Model Validity

Follow-up analyses tested the validity of the LIWC model by seizing the available dialogue act annotations available in the Map task corpus (presented in Table 2.2). The lexical count metrics of LIWC were compared with the dialogue act annotations, which were used as a gold-standard for Track 2 dialogue.[1] Analyses focused on the *Acknowledgement*, *Check*, and *Align* moves. *Acknowledgement* is explicit confirmation that a message has been heard and accepted. *Check* asks an interlocutor to confirm information that the speaker is unsure about. *Align* is a speaker's attempt to gain evidence of understanding from the interlocutor.

LIWC *Assent* and *Certain* content were examined through their identification of dialogue moves. The *Assent* word list identified 53% of the Acknowledgement moves, 7% of the Check moves, and 37% of the Align moves. Of all the utterances that contained one

[1] H. Clark argues that every contribution has a Track 2 element and I do not dispute this. Here analyses focused on acts that are predominantly Track 2 in nature.

or more words from the Assent list, 48.8% of those were one of the three Track 2 dialogue moves. The *Certain* word list identified 1% of the Acknowledgement moves, 3% of the Check moves, and 2% of the Align moves. Of all the utterances that contained one or more words from the Certainty list, 21% of those were one of the three Track 2 dialogue moves. In summary, there was validity to the LIWC model Track 2 dialogue, which was stronger for the Assent word list than the Certain word list.

Results: Correlations between Models

As a test of the models' construct validity, correlations examined the relationships between the alignment, coordination, and baseline models. As alignment and coordination are alternative accounts of grounding phenomenon, models representing each theory should not be related to each other. Correlations between alignment, coordination, and baseline recurrence metrics investigated how different these models were, and therefore tested if the recurrence models are measuring what they claim to measure (i.e., unique construct validity).

A number of unexpected relationships appeared. The pitch level and rhythm level correlation matrices were generally positive (Figure 6.1 and 6.2). Though correlations for pitch level recurrence rate (RR) were expected due to the recurrence rate normalization procedure, relationships appeared across all the recurrence metrics. At the other levels of analysis, there were positive relationships between coordination and baseline models for the morpheme level (Figure 6.3), word level (Figure 6.4), and syntax level (Figure 6.5). The relationships between coordination and baseline were most apparent on the negative diagonal, i.e., RR_Coord and RR_Base; DET_Coord and DET_Base, etc. Relationships appeared between alignment and coordination models at the morpheme, word, and syntax levels but these were reduced in comparison to the pitch and rhythm levels.

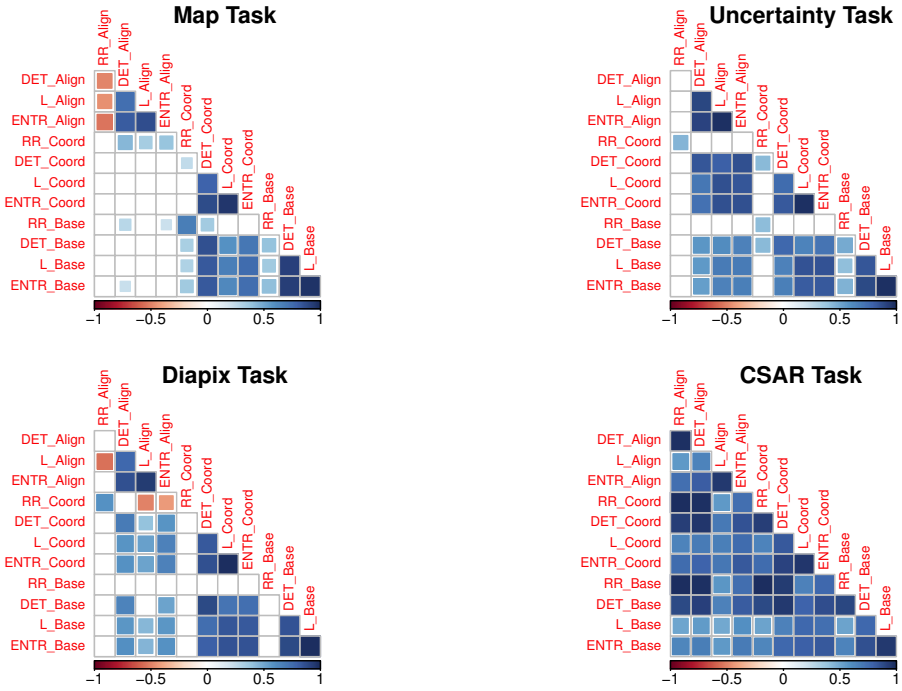

Figure 6.1: Correlation matrix for the pitch level for each task. Note: correlations with $p >$ .01 are omitted.

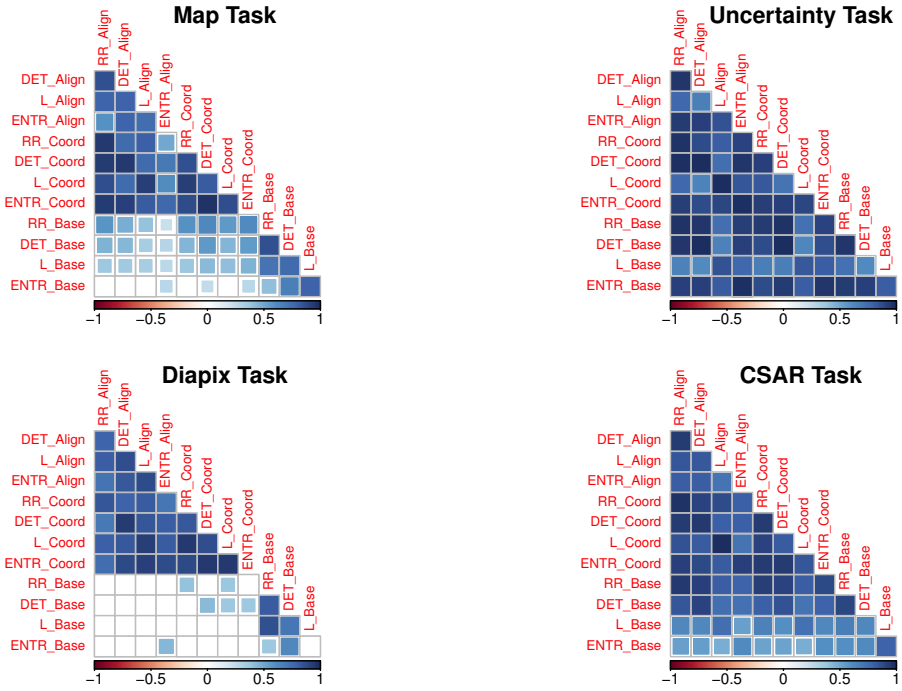

Figure 6.2: Correlation matrix for the rhythm level for each task. Note: correlations with p $>$ .01 are omitted.

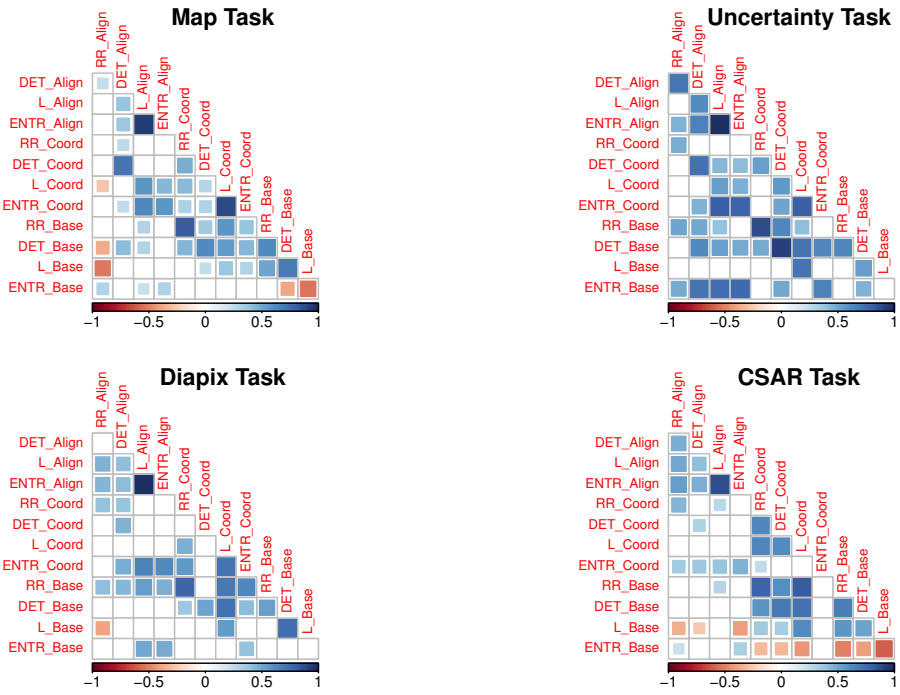

Figure 6.3: Correlation matrix for the morpheme level for each task. Note: correlations with $p > .01$ are omitted.

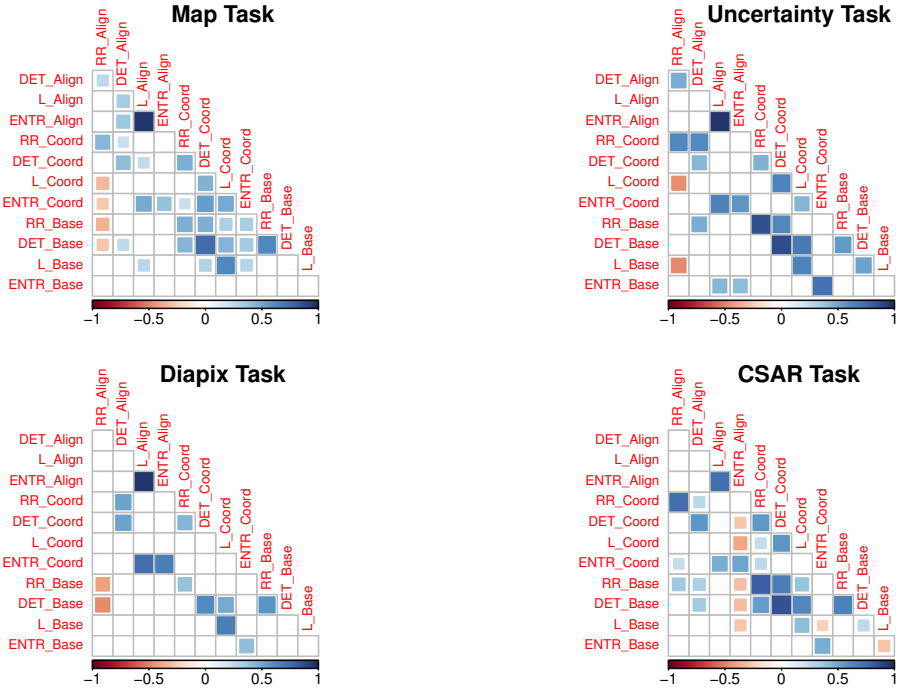

Figure 6.4: Correlation matrix for the word level for each task. Note: correlations with $p > .01$ are omitted.

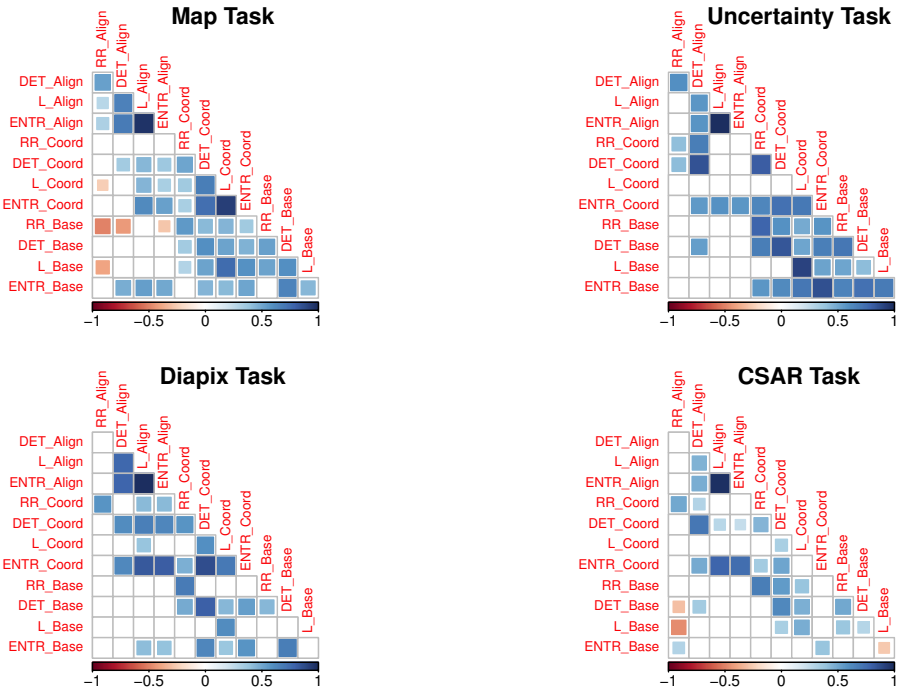

Figure 6.5: Correlation matrix for the syntax level for each task. Note: correlations with $p > .01$ are omitted.

Results: Predicting Accuracy vs. Time

These analyses used the accuracy and time performance measures to further examine the contents of the recurrence models. A valid model of grounding processes should predict both time and accuracy performance metrics. An additional accuracy metric was extracted from the Uncertainty task from the first submission, as described in Section 2.1.2. This provided two tasks with accuracy metrics (i.e., Map task and Uncertainty task) to have greater insight into the general ability of these recurrence models to predict task accuracy. In addition, a model of grounding should represent a time series's content rather than its length, so additional analyses examined the relationship between recurrence metrics and word count. Word count was calculated for each trial of each task and related to performance, particularly completion time measures, and to the recurrence models. Follow-on analyses synthesized shorter trials by removing sections from longer trials. Recurrence metrics from synthesized data were compared to the recurrence metrics from original data. If the synthesized recurrence metrics are distinct from the original recurrence metrics, we can conclude that the recurrence metrics are measuring some of the content and not solely a reflection of word count.

Results showed that the recurrence models explain less variance in task accuracy than competition time. These analyses also showed that the recurrence models' success in predicting task completion time is partly due to a strong relationship with word count.

7.1 Uncertainty Task Accuracy

These analyses investigated why prediction of accuracy measures (i.e., Map task path deviation) was statistically significant but less successful than completion time measures. On average, the coordination and baseline models predicted approximately 42% less of the variance in accuracy measures than completion time in the Map task. A correlation between the Map task path deviation score and completion time was not significant, r = -0.10, $t(126)$ = -1.09, p = 0.27. Therefore, prediction of completion time would not necessarily entail prediction of task accuracy.

Additional data extracted from the Uncertainty Elicitation task supported a second prediction of accuracy measures using the morpheme-, word- and syntax-level models (a subset of the levels analyzed above). This accuracy metric was not significantly correlated to the Uncertainty task final completion time, r = -0.10, $t(38)$ = -0.65, p = 0.52. The coordination model and baseline model were both significant at multiple levels of analysis (Figure 7.1). The baseline morpheme level model outperformed all the coordination models. Appendix B presents all model details. In comparison to the Uncertainty completion time analyses presented in Figure 4.2, variance accounted for in accuracy is approximately 43% less on average for the baseline and coordination models. In summary, for both the Map task and the Uncertainty task the recurrence models explain much less variance in task accuracy than they do in task completion time. Additionally, the pattern of results in the Uncertainty was different from other analyses in that the baseline models were better predictors of performance than coordination models.

7.2 Word Count

Analyses investigated the successful prediction of completion time through correlations with word count. Word count is strongly correlated to completion time, with longer com-

pletion times having more words as expected. Pearson's r values ranged from 0.58 to 0.94 across the tasks (Table 7.1). The Uncertainty Elicitation task had the strongest relationship ($r = 0.94$), followed by the Map task completion time ($r = 0.88$), the CSAR task ($r = 0.80$), and the Diapix task ($r = 0.58$). This ordering closely followed the ordering of the coordination recurrence model predictions: the Uncertainty task ($Adj\ R^2 = 0.76$), followed by the Map task completion time ($Adj\ R^2 = 0.63$), the CSAR task ($Adj\ R^2 = 0.45$), and the Diapix task ($Adj\ R^2 = 0.29$).

Table 7.1: Correlations between word count and task performance—See text for details. (***$p < .001$)

Task (Measure)	***r***	***t***	**df**	***p*-value**
Diapix (Time)	0.58***	4.86	46	$< .001$
Uncertainty (Time)	0.94***	16.64	38	$< .001$
Uncertainty (Accuracy)	0.26	1.65	38	0.11
CSAR (Time)	0.80***	13.84	108	$< .001$
Map (Time)	0.88***	20.94	126	$< .001$
Map (Accuracy)	-0.16	-1.90	126	0.06

In follow-on analyses, word count functioned as a mediator in a test of statistical mediation of the relationship between the word-level coordination model and task completion time. Table 7.2 summarizes the results, showing only Step 3 of the mediation analyses.[1] For the Diapix, Uncertainty, and Map tasks, word count completely mediated the relationship between the word-level coordination model and task completion time. For the CSAR task, word count partially mediated the relationship between the word-level coordination model and task completion time.

Last, an analysis synthesized data of different lengths from one trial from the Uncertainty Elicitation task. This trial had 2,139 words originally placing it in the 75% percentile of trials. It was shortened from 2000 words to 250 words in steps of 250. Each shortened

[1] Step 1, the relationship between word-level coordination model and task completion times, is shown in Chapter 4. Step 2, the relationship between the word-level coordination model and word count, is not shown though it was significant for each corpus.

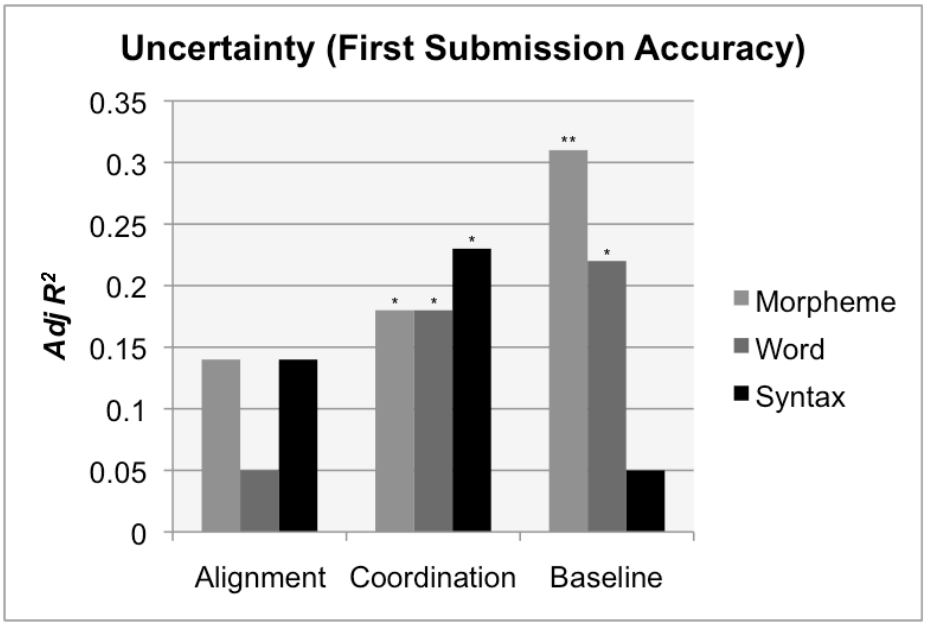

Figure 7.1: Overview of recurrence models prediction of first submission accuracy in the Uncertainty Elicitation task. Note: the ordinate range differs from the other plots. (* $p < .05$; ** $p < .01$)

Table 7.2: Summary of word count mediation tests for completion time measures. β and β' are the standardized regression coefficients without and with word count in the model, respectively. (** $p < .01$, *** $p < .001$).

Task	**Predictor**	β	β'
CSAR	RR	-0.22**	-0.03
	DET	-0.63***	+0.18**
	L	+0.01	-0.03
	ENTR	+0.13	+0.08
Uncertainty	RR	-0.34**	-0.06
	DET	-0.20	+0.05
	L	-0.72***	-0.09
	ENTR	+0.15	+0.07
Diapix	RR	-0.30***	+0.33
	DET	-0.49***	-0.08
	L	-0.59***	+0.15
	ENTR	+0.08	-0.08
Map Task	RR	-0.41***	-0.11
	DET	-0.23**	+0.04
	L	-0.56***	-0.04
	ENTR	+0.27***	0.02

trial was used to create a word-level coordination model, then the resulting recurrence metrics were compared to the recurrence metrics from the original trials.

Figures 7.2, 7.3, 7.4 and 7.5 show the recurrence rate, determinism, line length, and line entropy scores as a function of word count. For the original recurrence rate and line entropy metrics, there are weak relationships to word count, so the appearance of synthesized within the original data is not compelling. However, the original determinism and average line length metric data is clearly related to word count. Here the synthesized data follows the original closely.

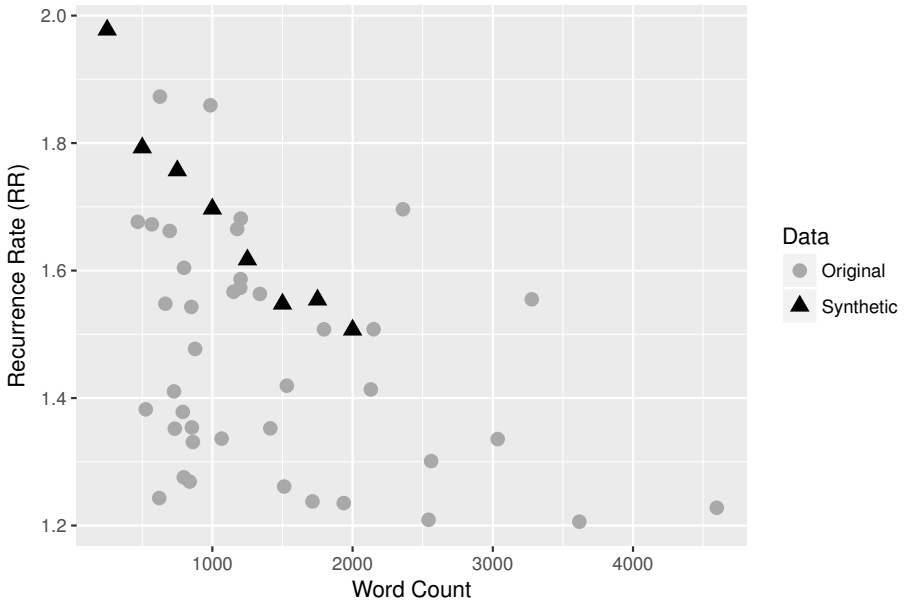

Figure 7.2: Recurrence Rate values (RR) as a function of word count. Synthesized data by shortening one trial (black) overlaid on the original data (gray).

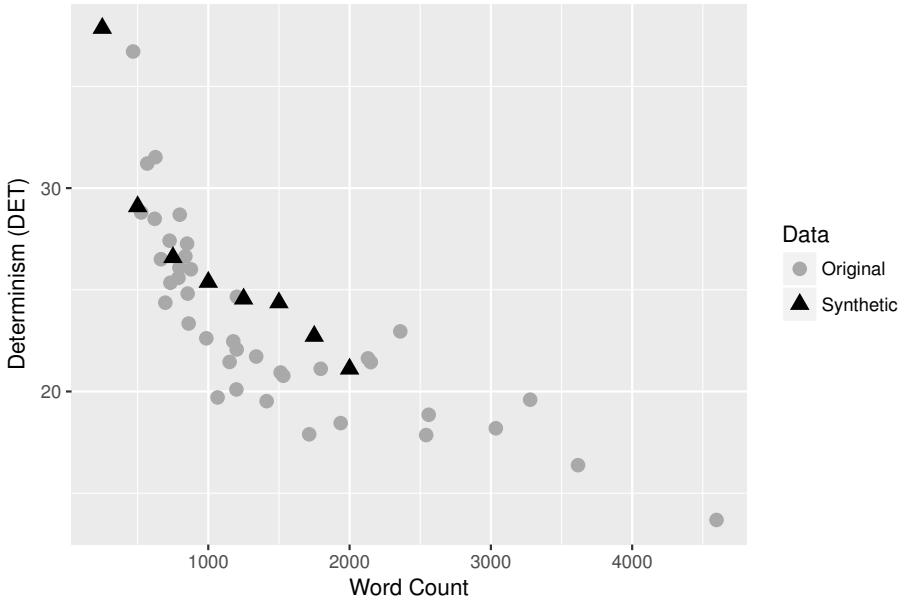

Figure 7.3: Determinism values (DET) as a function of word count. Synthesized data by shortening one trial (black) overlaid on the original data (gray).

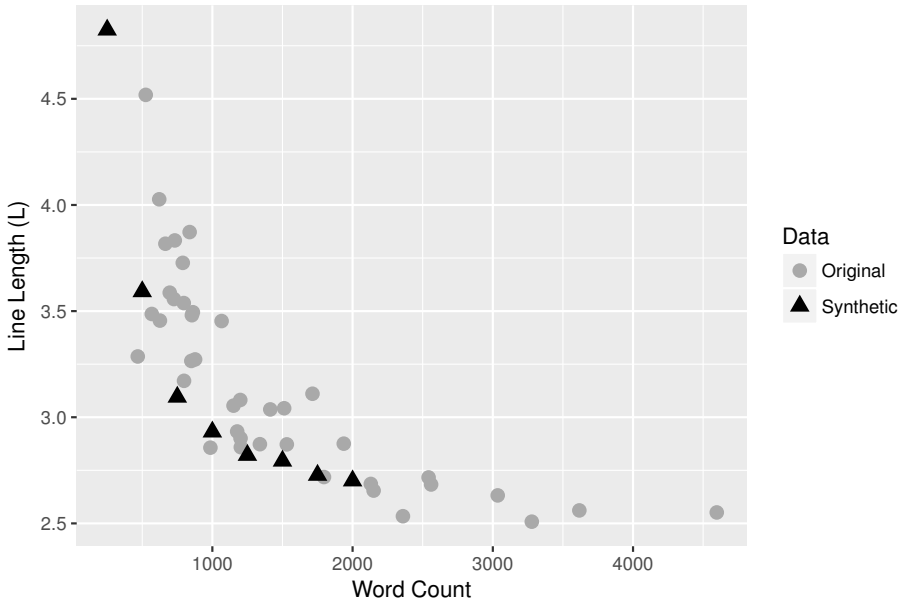

Figure 7.4: Average Line Length values (L) as a function of word count. Synthesized data by shortening one trial (black) overlaid on the original data (gray).

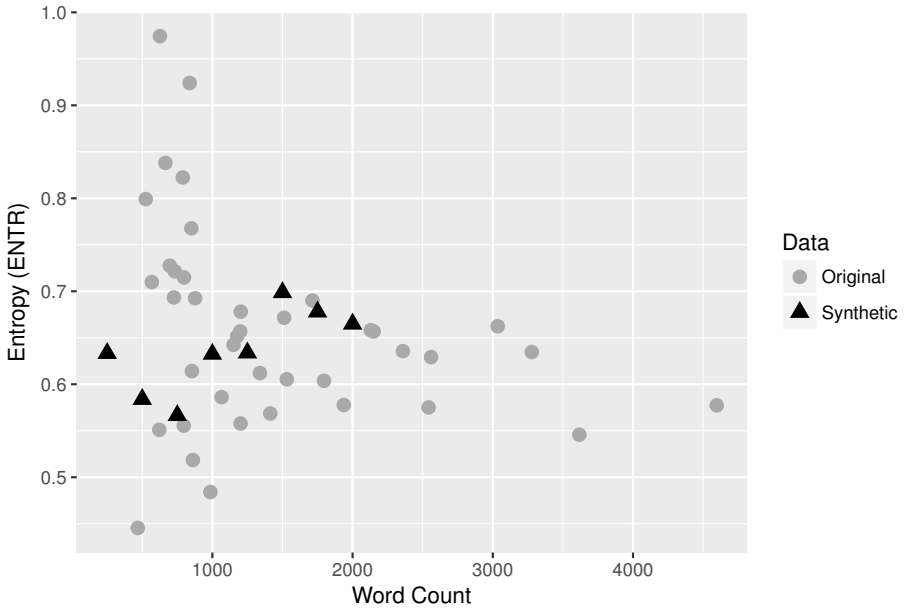

Figure 7.5: Line Entropy values (ENTR) as a function of word count. Synthesized data by shortening one trial (black) overlaid on the original data (gray).

Results: Additional Communication Metrics

As a final examination of the contents of the recurrence metrics, analyses tested their relationship to previously introduced methods of characterizing communication. These methods spanned measurement of grounding, workload, and articulation work. An additional purpose for these analyses was to focus on communication metrics that were less computationally demanding than recurrence calculations and therefore could potentially be utilized in a dialogue system and also in real-time communication monitoring and assessment. A manipulation check suggested the presence of articulation work through a relationship with swearing and negative emotion. These analyses were conducted on the CSAR task only, because the task may demand articulation work. Furthermore, the analysis used the word-level recurrence models because these models best predicted performance for the alignment model and coordination model in the CSAR task.

8.1 Grounding: Length of Installment

The average length of installment consisted of the median number of words per turn for each trial. Length of installment was not related to task completion times, $Adj\ R^2 = 0.00$, $F(1, 108) = 0.31$, $p = .58$.

Regression analyses treated recurrence metrics as predictors of the length of installment measure. The word-level alignment model was not significantly related to the length of installment, $Adj\ R^2 = 0.02$, $F(4, 113) = 1.62$, $p = .17$. The word-level coordination model and the word-level baseline model were significantly related to length of installment, $Adj\ R^2 = 0.23$, $F(4, 115) = 9.95$, $p < .001$, and $Adj\ R^2 = 0.20$, $F(4, 114) = 8.14$, $p < .001$, respectively.

Table 8.1: Regressions on average length of installment using word-level models. Regression coefficients (and standard error) are shown for each recurrence metric of each model.

	Dependent variable: Median Installment Length		
	Model Type		
	Alignment	**Coordination**	**Baseline**
RR	−2.255	−1.026	−0.099
	(3.780)	(0.815)	(0.329)
DET	0.083*	0.117***	0.102***
	(0.047)	(0.030)	(0.024)
L	2.235	0.075**	−0.022
	(2.268)	(0.029)	(0.025)
ENTR	0.145	4.393***	0.632
	(1.475)	(1.259)	(0.893)
Constant	−0.919	−0.691	−0.278
	(4.789)	(1.203)	(0.986)
Observations	118	120	119
R^2	0.054	0.257	0.222
Adjusted R^2	0.021	0.231	0.195
Residual Std. Error	2.643 (df = 113)	2.446 (df = 115)	2.473 (df = 114)
F Statistic	1.617 (df = 4; 113)	9.949*** (df = 4; 115)	8.135*** (df = 4; 114)
Note:			*p<0.1; **p<0.05; ***p<0.01

8.2 Articulation Work: Pronouns

Articulation work often addresses the dependency relationships in a team, such as shifts from independent to collective task processes, therefore these analyses focused on singular and plural pronouns following Khawaja et al. (2012, 2014). Specifically, LIWC lexicons for singular pronouns *I* and *he/she* and plural pronouns *we* and *they* were calculated. LIWC analyzed the presence of each word or word list by counting occurrences in each trial then normalizing the counts by the word count for that trial.

The singular pronouns served as predictors in one regression on completion times and the plural pronouns served as predictors in a second regression on completion times. Neither model was significant, $Adj\ R^2 = 0.00$, $F(2,\ 107) = 0.01$, $p = 0.90$, and $Adj\ R^2 = 0.01$, $F(2,\ 107) = 1.67$, $p = .19$.

8.3 Articulation Work: Swearing and Negative Emotion

A manipulation check tested the presence of articulation work in the CSAR task. The LIWC lexicons for *Swear* and *Negative Emotion* were used as coarse measures of task demands that were likely to provoke articulation work, such as plans being interrupted due to enemy combatants. Examples of the *Swear* word list included: crap, dang, heck, and sucks. Examples of the *Negative Emotion* word list included: angry, dumb, lame, sad. LIWC analyzed the presence of each word list by counting occurrences in each trial then normalizing the counts by the word count for that trial.

In separate regressions, each word list was regressed on task completion times. The *Swear* word list was a significant predictor of performance, $Adj\ R^2 = 0.16$, $F(1,\ 108) = 22.04$, $p < .001$. The *Negative Emotion* word list was a significant predictor of performance as well, $Adj\ R^2 = 0.03$, $F(1,\ 108) = 4.83$, $p < .05$. As both swearing and negative emotion increased, there tended to be longer completion times.

Additional analyses tested for relationships between articulation work and grounding processes. Each recurrence model attempted to predict *Swear* and *Negative Emotion* lists, using the word-level models. Table 8.2 shows the results for the *Swear* list. Alignment, coordination and baseline models were all significant (all $p < .01$). The coordination model explained more variance (18%) in swearing than alignment or baseline (12% & 10%, respectively). Table 8.3 shows the results for the *Negative Emotion* list. Alignment, coordination and baseline models were all significant (all $p < .01$). The baseline and coordination model explained more variance in negative emotion (11% & 10%, respectively) than the alignment model (8%).

Table 8.2: Regressions on Swearing. Regression coefficients (and standard error) are shown for each recurrence metric of each model.

	Dependent variable: Swearing		
	Model Type		
	Alignment	**Coordination**	**Baseline**
RR	-0.520^{***}	-0.087^{**}	-0.044^{**}
	(0.188)	(0.042)	(0.018)
DET	-0.007^{***}	-0.004^{**}	-0.001
	(0.002)	(0.002)	(0.001)
L	-0.213^{*}	-0.0003	-0.0002
	(0.113)	(0.002)	(0.001)
ENTR	0.170^{**}	-0.041	-0.016
	(0.073)	(0.065)	(0.048)
Constant	0.727^{***}	0.383^{***}	0.237^{***}
	(0.238)	(0.062)	(0.053)
Observations	118	120	119
R^2	0.153	0.211	0.129
Adjusted R^2	0.124	0.183	0.099
Residual Std. Error	0.131 (df = 113)	0.126 (df = 115)	0.133 (df = 114)
F Statistic	5.123^{***} (df = 4; 113)	7.671^{***} (df = 4; 115)	4.226^{***} (df = 4; 114)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 8.3: Regressions on Negative Emotion. Regression coefficients (and standard error) are shown for each recurrence metric of each model.

	Dependent variable: Negative Emotion		
	Model Type		
	Alignment	**Coordination**	**Baseline**
RR	−0.950	−0.304**	−0.155***
	(0.588)	(0.135)	(0.054)
DET	−0.022***	−0.006	−0.001
	(0.007)	(0.005)	(0.004)
L	−0.558	−0.002	−0.003
	(0.352)	(0.005)	(0.004)
ENTR	0.476**	−0.066	0.029
	(0.229)	(0.208)	(0.146)
Constant	2.022***	1.171***	0.855***
	(0.744)	(0.199)	(0.161)
Observations	118	120	119
R^2	0.110	0.134	0.144
Adjusted R^2	0.079	0.104	0.114
Residual Std. Error	0.411 (df = 113)	0.405 (df = 115)	0.403 (df = 114)
F Statistic	3.498*** (df = 4; 113)	4.439*** (df = 4; 115)	4.800*** (df = 4; 114)
Note:			*p<0.1; **p<0.05; ***p<0.01

Discussion

9.1 Summary of Results

This research examined communication behavior in four different team tasks to investigate the contribution of conversational grounding processes to task performance. The primary research question centered on the proper model for grounding. Both alignment and coordination models of grounding were successful in predicting task performance, showing a clear connection between communication behavior and task outcomes. However, the pattern of results clearly supported the coordination model and revealed deficiencies of the alignment model. This pattern of results was supported across the tasks and across multiple levels of linguistic analysis. A variant of coordination, strategic design, was tested using statistical mediation analysis with a representation of Track 2 dialogue. The analysis partially supported audience design as a viable variant of the coordination model. However, the models of grounding performed differently across the various tasks, and sizeable differences in the performance outcome relationship occurred between completion time and accuracy metrics. Differences emerged across levels of linguistic analysis as well, revealed by contrasting the consistent performance of the lexical levels (word and morpheme) with the inconsistent performance of the pitch and rhythm levels.

Additional tests investigated the construct validity of the models. High observed correlations between alignment and coordination models at the pitch and rhythm levels indicated poor construct validity at these levels. In contrast, construct validity was reinforced

for the word-level coordination model by the relationship between the word-level coordination model and the length of installments. Finally, task complexity arose as an important element because of the requirement for the articulation work that entails task management dialogues. Analyses of swearing and negative emotion indicated articulation work was provoked by CSAR task characteristics and, importantly, this was related to conversational grounding models.

9.2 Common Ground and Task Performance

Common ground and understanding are challenging constructs to measure. They are not directly observable or quantifiable, effect size is incalculable because the variance is unknown, and the notion of a grounding criterion (discussed in Clark and Schaefer, 1989) suggests that understanding is always contextualized to current purposes. Furthermore, any indirect measures have limited sensitivity as well as coarse time resolution. The foundational inference for the current research hinges on a relationship between task performance and common ground. For a task that requires common ground to succeed, the task performance metric captures some amount of the grounding process. Therefore, measurements of communication that can predict task performance are representing a component of common ground and the grounding process.

9.2.1 Communication Processes Predict Task Performance

Communication processes represented by quantitative recurrence models successfully predicted task performance. This impressive relationship occured for spontaneous speech in each of the four different tasks and for both the completion time *and* the task accuracy performance measures. To the author's knowledge, no single study has predicted performance in *four different tasks simultaneously* with the same models of communication behavior.

Clearly, recurrence models capture a general feature of dialogue.

9.2.1.1 Prior work relating common ground to performance

Providing a quantitative relationship between common ground and performance is unique. Research on computer-supported cooperative work addresses the relationship between communication and task outcomes, though with qualitative methods (e.g., Carroll et al., 2003). Human factors researchers identify the influence of communication, (though not specifically common ground), on task performance, also commonly referred to as team effectiveness. However, the most popular analysis method merely bifurcates teams into good or bad categories. The split is sometimes based on median performance (e.g., Gervits et al., 2016) or based on the tails of the performance distribution (e.g., Fischer et al., 2007). The good-bad bifurcation method has provided some insight into successful communication. Butchibabu et al. (2016) found that compared to bad teams, good teams use a communication strategy of anticipating teammates' information needs. Bad teams mostly waited for teammates to request information. Burtscher et al. (2010) found that in response to nonroutine events, good teams increase "task management communication" more than bad teams. Task management communication included planning, task distribution, requesting/offering assistance, which are components of articulation work.

Recently, some researchers have used a good-bad bifurcation to investigate common ground phenomena. Though not using the term common ground, Fischer et al. (2007) found that good teams elaborated more compared to bad teams, which would be expected to contribute to increased understanding. They also found that bad teams failed to respond to interlocutors (i.e., failed to 'close the loop') more often than good teams. Gervits et al. (2016) used a distributed search task to examine many differences in communication between good and bad teams, including disfluencies and grounding strategies. Compared to bad teams, good teams increased their disfluencies but engaged in more self-repair. They argued that this indicated a sensitivity to their addressees' perspectives by anticipating am-

biguity and offering unsolicited clarification. They also found that good teams had a higher rate of *Check* and *Ready* dialogue moves that serve to check the understanding of the addressee and explicitly state the ability to initiate a new sub-task, respectively.

Although the good-bad bifurcation is informative for preliminary insights into what makes good teams good, it lacks the sensitivity to compare between the multiple behaviors that appear in good teams. While mechanism is theoretically important, application efforts depend on identifying those aspects of communication that provide the largest contribution to team performance to guide the design of both dialogue systems and human communication interventions. The needed sensitivity only comes with quantitative predictions of performance outcomes.

Other research has related communication, and specifically common ground, to quantitative performance in a single task. The classic tangram task (e.g., Clark and Wilkes-Gibbs, 1986) clearly exemplifies how collaborative referential expressions are developed, reused, and shortened over successive trials. The shortening leads to gains in efficiency and as a result, trial completion time decreases dramatically. An extension of the tangram paradigm provides additional support for the common ground-performance relationship. Weber and Camerer (2003) perturbed the partners mid-way through the experiment and found that this disrupted the previously established common ground and led to increased trial times. Other work has shown how the addition of visual information, such as a shared workspace, can increase common ground and result in performance benefits like reduced task completion time (Clark and Krych, 2004; Gergle et al., 2004). Unfortunately, none of the above studies reported effect size or variance explained, so it's difficult to assess the size of the relationship between common ground and performance.

Two studies summarized in the introduction were able to quantify the relationship between communication and performance. Kiekel et al. (2002) examined team unmanned aerial vehicle task performance and its relationship to the dialogue content measured by Latent Semantic Analysis (LSA), which was compared to trials with known task performance.

They were able to explain 39% of the variance in their composite performance metric. Yee et al. (2017) studied a team Tower of Hanoi task and found that more closing-the-loop communication was related to faster task completion times (the standardized regression coefficient reported was .41). Interestingly, closing-the-loop communication mediated the relationship between social cue utilization and task performance, where an aspect of social cue utilization could include non-verbal signals of understanding. Yee et al.'s results and the results of Fischer et al. (2007) described above correspond with the current research's emphasis on coordination—complementarity of information and action provide for successful communication. The adjacency pairs of initiation and closure create common ground, not alignment/similarity.

Reitter and Moore (2014) is directly comparable to the current research. As mentioned in the introduction, they predicted path deviation performance in the HCRC Map Task corpus, which the current research did as well. They trained a support vector machine using a combination of lexical and syntactic features measuring long-term adaptation. The adaptation features were on the time scale of minutes, and therefore did not represent classic priming-based alignment, which has a much shorter time scale.[1] Instead, this long-scale alignment might reflect implicit learning. The current methods did not measure change over the course of a trial, which would be analogous to Reitter and Moore's study of long-scale alignment between the beginning and end of a trial. The current work did examine the learning on a larger timescale, over multiple trials, and found that this had a mixed affect on performance (i.e., only some teams for some tasks showed significant learning). In addition, when learning was accounted for, the recurrence models' performance predictions were only slightly affected. Reitter and Moore were able to predict 17% of the variance in task performance using a combination of lexical and syntactic features. In the current

[1]Reitter and Moore suggest Pickering and Garrod do not specify the time scale of alignment, "Pickering and Garrod (2004) do not detail the longevity of the priming effects supporting alignment. It is unclear whether alignment is due to the automatic, classical priming effect, or whether it is based on a long-term effect that is possibly related to implicit learning" (p. 10-11). However, it is difficult not to take a plain reading of 'priming' for each time it is used.

research, the syntax level coordination model accounted for 9% of performance, the word level coordination model accounted for 5% and the rhythm level accounted for 6%. The rhythm level alignment model also account for 8% of the variance in performance. The Reitter and Moore value reported was R^2 for three predictors whereas the values in the current research were *Adjusted* R^2 using four predictors.

9.2.2 Coordination Bests Alignment

The results clearly supported the coordination model over the alignment model. The coordination models were significant predictors of performance more often than the alignment models and this pattern was consistent across tasks. In addition, the coordination models explained more variance in performance than the alignment models for *both* completion time and task accuracy measures. The pattern of findings for alignment and coordination was similar to Fusaroli and Tylén (2016). Fusaroli had a symmetric task and here I found similar results for symmetric tasks: Uncertainty, Diapix, and CSAR. In addition, I also found a similar result for an asymmetric task (Map Task), which the recurrence models have not been applied to yet.

The superiority of coordination corresponds with other recent developments in common ground research. Research suggests that interlocutors diverge over time and become more complementary rather than become more similar (Mills, 2014). These changes occur across multiple trials spread out over 90 minutes. A likely explanation for divergence over time is that interlocutors provide new content in their contributions rather than repeating content from previous contributions (Tenbrink et al., 2008). The excerpt in Table 2.4 contained an instance of new complementary content when Speaker A said "*I see it and there's like a stop sign*." The additional mention of the stop sign provided very strong evidence that he had identified the house that Speaker B was describing. Assertions like these have a Track 2 function because they explicitly signal the understanding Speaker A has and invite Speaker B to object if necessary.

Returning to Butchibabu et al. (2016) discussed above, they found that good teams *anticipated* the information needs of their teammates. A trend of anticipatory communication increasing over time also appears in Convertino et al. (2008, 2009) and anticipatory communication is one common outcome of cross-training (Salas et al., 2008). Importantly, anticipation involves two aspects that are indicative of audience design. First, there is a recognition of Theory of Mind—my teammates do not have access to all the information I have access to. Second, there must be some knowledge and representation of what your teammates need and when they need it. Other work has also shown that dialogue contributions often reflect different perspectives and interlocutors appear to maintain and keep track of multiple perspectives at the same time (Brennan et al., 2013). In addition, participants take their partners' perspectives despite the increase in cognitive demand required to do so, which may be due to a default attitude of collaboration (Duran et al., 2011).

The current results also showed the deficiency of alignment as a model of common ground. Perhaps the failure of alignment lies in a failed claim of propagation across multiple linguistic levels. Pickering and Garrod (2004) argued that alignment at one level leads to alignment at other levels eventually including the situation model. They review many studies detailing low-level alignment (e.g., phonetic convergence) and mid-level alignment (e.g., lexical entrainment) and they are undeniably well documented phenomena. Alignment at lower levels can lead to alignment at higher levels of linguistic complexity, and eventually the level of the situational model that encompasses the shared understanding of common ground. Along these lines, Branigan et al. (2000) showed that alignment at the lexical level led to more alignment at the syntax level. The current research is unique in that it simultaneously tested alignment at 5 levels (Fusaroli and Tylén used three and most prior work uses two). For 3 of the 4 tasks studied, alignment models were significant at 2 or fewer levels of the 5 possible. Perhaps the relationship to performance was lacking because alignment had not appeared across more levels. (An exception was the Map Task completion time data, which found significance at 4 of 5 alignment levels and explained up

to 12% of the variance in performance, see Figure 4.5).

The current results on limited alignment success can be compared to Reitter and Moore (2014), discussed above, which found that short-term alignment did not predict task performance. Their study used the Map Task corpus path deviation metric and analyzed the syntax level. The current study's findings at the syntax level also showed that alignment could not predict performance. The current study has added to the results of Reitter and Moore (2014) in a number of ways. Here, in addition to syntax, the alignment models failed to predict performance at the pitch, morpheme and word level. One level was significant, however. At the rhythm-level the alignment model predicted performance and this relationship was maintained after controlling for learning effects ($AdjR^2 = 0.07$).[2]

9.2.3 Which Variant of Coordination

Coordination has been represented by two different theories: interpersonal synergy and audience design. The difference centers on the intentionality of the communication (in both production and comprehension). Audience design argues that coordination is the result of an intentional act that relies on Theory of Mind. Interpersonal synergy argues that coordination can be an unintentional process that emerges from the interaction. Statistical mediation used a LIWC model of Track 2 dialogue to test if the variant of coordination was interpersonal synergy or audience design. There was partial support for audience design. The coordination recurrence model completely mediated the relationship between Track 2 dialogue and performance in the Uncertainty task. The coordination recurrence model partially mediated the relationship between Track 2 dialogue and performance for the Map task completion time metric. By contrast, the CSAR and Diapix tasks and the Map task path deviation measure had no mediation. In all three cases, this was due to a failure of the LIWC model of Track 2 dialogue to relate to performance. The common ground that results from Track 2 dialogue clearly plays a role in performance, as discussed above, so

[2]The rhythm level was not analyzed in Reitter and Moore.

this result could be indicative of a methodological issue, which will be discussed further below.

Another piece of evidence for audience design comes from the analysis of installment length. Prior research has identified that speakers can use a communication strategy for a complex or important message—break it up into separate installments (Clark and Schaefer, 1987). In the current study, the length of installment was related to the coordination and baseline recurrence models but not related to the alignment model or to task performance. This finding is consistent with the notion of audience design because speakers would be considerate of the effort and likely failure of a long message and would choose to make multiple installments.

A sequence of influential studies on referential communication has challenged when and how Theory of Mind is used in interpretation (Horton and Keysar, 1996; Keysar and Horton, 1998). Their research used a workspace that contained shared items seen by both speaker and addressee. There were also privileged items that were only visible to the addressee. The speaker was a confederate following a script. The speaker references one of the shared items, but findings indicated that addressees often selected an item that was in the privileged ground. In other words, addressees knew the item was not visible to the speaker and they still selected it. Keysar's interpretation was that the initial comprehension is egocentric—Theory of Mind considerations occur late, in a revisionist process of sorts. Recent literature, although preliminary, suggests that a methodological choice has influenced the results. Hawkins and Goodman (2016) replicated Keysar's findings and collected an additional modification with unscripted references. The performance of the unscripted pairs was better than the scripted pairs and Keysar's data. Moreover, the unscripted references showed that speakers often over-specified, presumably because of the knowledge of the existence of privileged ground even with ignorance of the privileged ground's contents. This suggests that speakers took into account Theory of Mind in designing their references, and that addressees might rely on the cooperative references that

were produced. In the scripted cases that matched items in shared and privileged ground, addressees had a representation of the speaker that was violated by the scripted reference and likely led to more errors.

There is an important distinction about what is coordinated, whether it is content or process (for a review, see Mills, 2014). Presumably these two different types of coordination could be governed by different mechanisms. Along these lines, Mills research suggests that team processes are not negotiated and instead, interaction routines evolve from the interaction. This description of process coordination appears to be a unintentional mechanism, which could implicate interpersonal synergy over audience design.

9.3 Common Ground: Measurement and Task

The current research addressed a number of methodological issues surrounding dialogue research, and specifically centered on how to measure common ground and the impact of the task.

9.3.1 Bridging Content to Quantitative Measures

Research quantifying the impact of communication processes on team performance is sparse. This is due to an inherent challenge—the characterization of communication that effectively bridges from content analysis to quantitative measures. Content analysis is typically qualitative and the change from qualitative to quantitative is non-trivial. Often, qualitative and quantitative are viewed as only complementary measurements that lead to converging evidence on a particular phenomenon. Successfully bridging the two goes further; it requires a synthesis. One aspect that makes synthesis challenging is the *construct* validity of the quantitative items, that it measures what it actually purports to measure. Construct validity will be a theme in the following discussion, particularly with regards to the Track 2

dialogue model, the pitch and rhythm levels of analysis, and the baseline recurrence model based on self-consistency.

9.3.2 The Track 2 Dialogue Model

Statistical mediation informed the composition of the recurrence-based coordination model. The statistical mediation analysis relied on Track 2 dialogue, which was modeled using LIWC. The LIWC model was advantageous for its simplicity—it used counts of words that are likely to appear in Track 2 dialogue. However, the LIWC model had borderline validity. When compared to dialogue act annotations for the Map task corpus, the LIWC model was related for the *Assent* list but not the *Certain* list. This presents some ambiguity regarding the mediation analysis. It is possible that the coordination is interpersonal synergy when not mediated by Track 2 dialogue or that coordination is strategic design and mediation failed because of methodological problems with the Track 2 dialogue model implemented in LIWC.

9.3.3 Recurrence Analysis of Communication

The current research used models of common ground based on recurrence quantification analysis. A number of findings related to this nascent methodology, such as under what situations these models succeed or fail and how they can be improved.

9.3.3.1 Predicting Accuracy

Considering the predictions of different types of performance (presented in Section 4.1), it appeared that the recurrence models performed well on symmetric dialogue tasks but did not perform well on the asymmetric Map Task. Yet follow-up analysis indicated that the accuracy metric was problematic, and not the asymmetric dialogue setting. The models predicted performance in the Uncertainty task first submission measure that was extracted

from the data. For this task, the models explained more variance in performance than the Map Task but less than the Uncertainty completion time measure. Communication processes have a smaller role (but still significant) in task accuracy when compared to the role they play in task completion time.

We can conclude that asymmetry is not an inherent problem, but could present challenges approaching the limit where one speaker dominates the dialogue. For the alignment model, the points of recurrence will be limited severely by the less frequent speaker. For the coordination model and baseline models, the differentiation between these two models will be eroded because the dialogue will be one speaker primarily.

9.3.3.2 The Baseline Model

The baseline model was created to reflect self-consistency, which was contrasted with alignment and coordination. In the instance of alignment, self-consistency was expected to be low as the speakers increased their imitation of each other (i.e., alignment) over time. In the instance of coordination, the behavior of each speaker was expected to be similarly influenced by the other speaker due to the co-construction of the dialogue.

Yet the baseline model performed unexpectedly well in most tasks. For most tasks and levels of analysis, the baseline model explained more variance in performance than the alignment model. Moreover, in the Uncertainty task the baseline model explained almost as much variance in performance as the coordination model. This was an unexpected result and a departure from the findings in Fusaroli and Tylén (2016). In their work, the baseline model was not significantly related to performance at any level of analysis.

One possible explanation for the current research's findings is that the baseline model captures more than the self-consistency it was intended to capture. If communication is truly coordinative in nature, then a baseline model that only uses half of the dialogue (the time series from one speaker) will contain information about the other speaker's contributions. By analogy, when listening in on a friend's phone conversation it is often possible

to hear only your friend's contributions and infer a large amount of the inaudible contributions from the caller. This interpretation is reinforced by the strong positive correlations between baseline and coordination recurrence metrics for most tasks and levels of analysis. A similar finding appeared in Healey et al. (2014), where conversational participants were unlikely to repeat their own syntactic structures because of Gricean notions of relevance—they were responding appropriately to their interlocutors.

9.3.3.3 The Alignment Model

The measurement of alignment captured by the recurrence model is similar to the general alignment of Healey et al. (2014). Their metric examined the syntactic structures shared by the dyad, normalized by the total number of syntactic structures present in the dialogue. As previously mentioned in the Section 1.3.3, these general alignment measurements differ from the controlled laboratory tasks that typically focus on one pair of syntactic structures (e.g., the prepositional object structure and the double object structure). The measurement of syntactic alignment in natural conversational corpora is an important methodological advancement, but it is possible that global measurement techniques compare structures that are perhaps inappropriate to compare, and thereby underestimate the alignment present. For instance, an alternative method of measuring syntactic alignment introduced by Moscoso del Prado Martín and Du Bois (2015) found evidence of alignment in a conversation corpus. The differences between these proposed methods that explain the differences in findings have not been explored and are not simple to investigate. Recent methodological research in this area has increased in sophistication (e.g., Boghrati et al., 2017), further complicating comparisons.

9.3.3.4 The Construct Validity of the Pitch and Rhythm Levels

The current study identified problems with the continuous recurrence quantification analysis that examined the pitch and rhythm levels. Within each of the tasks analyzed, the

rhythm level tended to have similar relationship to performance for the alignment, coordination and baseline model. For instance, in the Uncertainty task the rhythm level models explained approximately 64% of the variance in performance for alignment, coordination and baseline. In the CSAR task, the rhythm level models explained approximately 5% of the variance in performance for alignment, coordination and baseline. This is likely due to the similarity of predictors; there were large positive correlations between the alignment, coordination, and baseline models for the pitch level (Figure 6.1) and the rhythm level (Figure 6.2). These problems indicate construct validity issues.

A possible cause of the construct validity issue is the normalization of pitch information. For the alignment phenomenon of pitch convergence, researchers typically measure the voice F0 of each speaker and how those come together overtime (e.g., Levitan and Hirschberg, 2011). The normalization procedure takes two speakers and gives them a similar mean, which would show them diverging instead of coming together. For instance, consider if Speaker A had an F0 of 150 Hz and Speaker B had an F0 of 175 Hz and they perfectly converge on an F0 of 162.5 Hz for the last 2 minutes of the trial. The F0 information would be normalized such that the mean of each F0 time series is 0. Afterwards, Speaker B's final 2 minutes would be below 0 and Speaker A's final two minutes would be above 0, which would not reflect the convergence as recurrence (depending somewhat on the radius that is used to define the neighborhood). Section 9.5 on future work will return to this problem and suggest some areas for improvement.

9.3.3.5 The Relationship between Coordination Model and Alignment Model

The relationship between the coordination model and the alignment model is complex. From a preliminary examination, it appears that the coordination model contains the alignment model because the recurrence of the whole dialogue necessarily includes the recurrence of Speaker A with Speaker B. This has a direct implication for the interpretation of the current findings, because if coordination includes alignment, we would expect coordination

to predict performance better than alignment. However, the correlation analyses between models at the morpheme, word, and syntax level did not show a strong positive relationship (Figures 6.3, 6.4 & 6.5). Recurrence quantification analysis (RQA) and cross recurrence quantification analysis (CRQA) are non-linear analysis techniques. Adding more points of recurrence to the recurrence plot has a non-linear affect on the metrics calculated from the plot and this could have accounted for the lack of relationship between the resulting RR, DET, L and ENTR metrics in the alignment and coordination models.

9.3.3.6 Non-stationarity

Upon finding a correlation between the word-level correlation model and trial word count, a test synthesized different length transcripts from one trial to investigate if recurrence metrics were affected by length alone with maximally similar content. The synthesized data showed a strong relationship between word count and the DET and L recurrence metrics (Figures 7.3 & 7.4). The relationship between word count and recurrence metrics can arise from 'measurement non-stationarity' (Rieke et al., 2004), which is due to insufficient observation time. In other words, the time span measured was shorter than the dynamics of the system such that the measured characteristics of the system change as observation time is increased. This situation may not be rare.

> "*There are many processes which are formally stationary when the limit of infinitely long observation times can be taken but which behave effectively like non-stationary processes when studied over finite times.*" (Kantz and Schreiber, 2004 , p. 14).

Additionally, human systems can exhibit *metastability*, complex dynamics with multiple regions of local stability and occasional transitions between regions (Gorman et al., 2017). Any observation time across a transition between regions could appear non-stationary. Particularly problematic, it is difficult to know the system dynamics prior to observation, to assure sufficient observation duration.

A natural following question is, what function could non-stationarity perform? One possibility is that the flexibility of language is one of its most powerful attributes. The law of requisite variety (Ashby, 1958) suggests that for stability any control system needs more variability than the process it is controlling. It follows that the flexibility and adaptability of language provides teams affordances to maintain stable control of a myriad of processes. Yet there is a balance between the flexibility consistent with the law of requisite variety and the inherently joint process of common ground. One speaker cannot depart too far from convention, or change too much or too fast, otherwise the other interlocutors would not be able to keep up.[3]

9.3.3.7 Truncating Completion Time

As just mentioned, the recurrence models were more successful at predicting task completion time than task accuracy. A standard practice is to analyze task completion time while accounting for accuracy, typically by holding accuracy constant or analyzing only a subset of trials that were correct. However, the CSAR task truncated completion times after 10 minutes if the rendezvous task was not accomplished. Though this was infrequent (5 of 120 trials), it would be problematic if it were more prevalent because it provides an artificially imposed cut-off of the task that assigns a completion time value when it was never completed. Future research should carefully consider how to handle these trials. Perhaps they should be analyzed separately from the correct data, in line with the standard practice. It's possible that the communication behavior in these truncated trials would provide much insight into failures to coordinate—what the team said that led to their poor performance. Furthermore, if these trials had been allowed to continue, they may have also provided examples of how teams recover from poor communication.

[3]Readers may find Bartel-Radic and Lesca (2011)'s related discussion of law of requisite variety and intercultural teams interesting.

9.3.4 Task Characteristics are Crucial

Although I demonstrated support for coordination across four different tasks, task characteristics remain crucial. Dialogue is a situated process. The current research underscored how the task plays a large role in the phenomenon observed and, as a result, the theory that can be developed. The vast majority of tasks in the common ground literature are solely communication tasks—there are no other sources of variance apart from the communication. Therefore, it was unclear if communication processes could still affect task performance in the presence of other contributors to performance. The current research began to address this through the tasks selected, and in particular the CSAR task. The CSAR task had many other contributions to task performance: the map was not known and could be obstructed, the enemy forces were mobile and armed. As will be discussed more below, the task aspects influenced the content of the dialogue by articulation work discussions, which is absent from communication-only tasks.

9.3.4.1 Articulation Work

A special type of communication complexity relates to what speakers talk about. The topics of dialogue in the current research included the task management activities involved in supporting articulation work, which is prevalent in human team communications. The articulation work topic is absent from most referential communication literature because of the prevalent use of simple tasks. It appeared in the current research because of the task characteristics in the CSAR task. The CSAR task exercised only a subset of all the different varieties of articulation work, but it did entail the task management issue of introducing new goals and short-term projects into the larger task. The 'to-be-rescued' did not have a weapon and could not defend his/herself from the enemy forces. In many instances the to-be-rescued communicated that s/he is being shot at and needs to change course to evade and hide. When this short-term project was complete, s/he could resume the primary goal of rendezvous. Moreover, the changes to the goals of one partner are not independent, they

have implications for the other partner. The to-be-rescueds could not defend themselves, so they introduced new goals for the rescuers by requesting the rescuers come and return fire on the enemy forces. Furthermore, this request only makes sense in light of a shared understanding (i.e., common ground) that one participant is armed and one is not, which is not explicitly stated in the goal communications.

The introduction of a new goal is a rare aspect of communication research tasks, so how teams discuss goals is an impoverished area of research. Recent research with similarly complex tasks provides a notable exception and underscores the notion of goals. Gervits et al. (2016) had a distributed search task that introduced a new goal 5 minutes into the trial. This goal was a surprise to participants, but reflective of the perturbations that are frequently experienced in situated task dialogue. Also, the goal was communicated to only one of the participants, who then had to describe it to his/her teammate. They found that goal communication was different between good and bad teams and poor conveyance of goals leads to poor task performance, as would be expected.

These instances of goal introduction were prompted by perturbations (i.e., depart from the routine flow of events) in the task. As previously discussed above, research has shown that good teams responded to nonroutine events with increased task management communications (Burtscher et al., 2010). Therefore, the capacity for articulation work is particularly critical for dialogue systems that are in domains with any probability of perturbation.

Even in static task settings without perturbations and introduction of new goals, articulation work can still be a critical piece of team communication. Convertino et al. (2008, 2009) argue that work in computer-mediated communication has been stinted because of a focus on grounding the task content and neglect of articulation work. Their work studied dialogue on task processes in an emergency management planning task that teams repeated three times. Early on in the experiment, the good teams had more task management conversation than bad teams. The good teams were successful in discussing their articulation work, as task management conversation decreased on later trials whereas the bad teams

required the same amount of task management throughout.

9.3.4.2 Task Complexity leads to Communication Complexity

The results from the current research are impressive because the tasks and the resulting dialogue were complex (and/or noisy). Some of the complexity stems from the stimuli used, such as the Uncertainty task's photographs of buildings, the CSAR task's immersive virtual scenes, and the Diapix task's cluttered pictures. For instance, the classic Chinese tangrams (Clark and Wilkes-Gibbs, 1986) were developed for their ambiguity, such that many different descriptions or conceptualizations could be used but this diversity is only apparent across pairs. Within pairs, the tendency is to only use one conceptualization, repeat it and truncate it. Other referential expression research that has departed from tangrams has used similarly simple stimuli, such as: filled and unfilled shapes (Horton and Keysar, 1996), solid or plaid squares (Gergle et al., 2013), and simple line drawings (Keysar et al., 2000). The stimulus complexity used in the current research was also greater than past research that used these recurrence models. Fusaroli and Tylén (2016) used a visual oddball detection task where stimuli were Gabor patches in a circular configuration that varied in visual contrast.

Some of the complexity stemmed from the joint activity that needed to be performed—the Map Task had participants draw free-hand routes, the CSAR task had participants navigate the virtual environment until a rendezvous. These are more complicated than referring to a single item (Horton and Keysar, 1996; Keysar et al., 2000), ordering stimuli (Clark and Wilkes-Gibbs, 1986), or discussing which interval (Fusaroli and Tylén, 2016). When the task is simple, it is unclear if the relationships found between communication processes and performance will generalize to complex tasks that have many additional factors contributing to performance.

9.3.4.3 A Caution on Causality

For communication-only tasks, the causal link from communication to performance is straight forward. Moving to complex tasks with many sources of variance beyond communication (e.g., the CSAR task) requires an important caution on causality. It is tempting to say that communication leads to task performance (or task difficulties), as in the communication-only tasks. However, it is also possible that the task affects communication and the communication merely identifies periods of time that are challenging to the team. In fact, this seems the appropriate interpretation of work like Grimm et al. (2017), where a perturbation in the environment preceded the change in communication (e.g., a fire began in the operating room and afterward the surgical team's communication changed). There may be some evidence of this in the current research, as the coordination model was related to word count. Difficult tasks could result in a long period of time that is filled with communication to accomplish the task.

Further complicating the issue of attribution, changes in communication can be *delayed*, such that one event prompts communications 5 minutes later, and communication can be *prolonged*, such that one event lasting 20 seconds prompts 10 minutes of communications (or the opposite—a 20 second exchange results in 10 minutes of joint action). Most circumstances do not have a fine level of temporal resolution to differentiate between the communication affecting task and task affecting communication, so it is important to avoid overstating causal directionality and acknowledge the ambiguity.

9.4 Applied Relevance

Many of the above results have implications for dialogue systems. The prediction of performance from dialogue behavior is widely useful in both human-human communication and human-machine communication settings. Systems that use alignment theory as inspiration have inherent limitations in supporting common ground. Alignment models were able

to predict performance in some instances, but to a small degree when compared with the coordination models. In addition, Branigan et al. (2010) suggests that some of the alignment observed between humans and computers is not due to alignment supporting common ground in dialogue. Rather, the observed alignment is due to expectations of low communication ability and a desire to achieve communication success. Humans overly adopt the dialogue system's lexicon because it is expected to facilitate effectiveness.

Some argue for the dynamical systems perspective as the foundation for dialogue (e.g., Fusaroli et al., 2014; Gorman et al., 2017) and therefore a natural implication is that dynamical systems notions should pervade dialogue technology development. Yet, it is not intuitive how to design a machine that could contribute to emergent stability or even exhibit metastability. The current research suggests that this alone doesn't inform how to represent interlocutors, represent successfully grounded material or specify what behaviors are necessary for successful common ground. The current work attempted connections to standard qualitative constructs such as Track 2 dialogue, and to standard quantitative items like length of installment and pronouns. Admittedly, this is a small dent in a very large area of research that should be explored further.

Within human-human team performance, there are direct applications to real-time communication monitoring. Models of communication processes such as these recurrence-based models could be employed in real-time communication monitoring. One challenge with any real-time monitoring application is detection latency, and dynamical systems measurements do not alleviate the issue (for a discussion see Gorman et al., 2012, which used analysis windows from 16-256 seconds). Recurrence quantification analysis requires a non-trivial number of samples for analysis. Marwan et al. (2007) suggests 1000 samples but perhaps 500 samples would suffice. The level of analysis can reduce latency as well. The morpheme level comprised of letter trigrams would reduce latency compared to the word and syntax levels. If the construct validity issues with the pitch and rhythm levels of analysis are resolved, they could provide the lowest latency.

9.4.1 Articulation Work

Perturbations are frequent in most work domains. Articulation work is an important topic of team dialogue as perturbations prompt articulation work. As a topic of dialogue, articulation would also be subject to the grounding process. The various ways participants could divvy up a task, and the many interdependencies between participants, is a complex and potentially confusing discussion. Ambiguity and misunderstanding are expected.

Swearing and *Negative Emotion* lexical items were measured by LIWC with the rationale that they arose from situations necessitating articulation work. Both these LIWC lexicons were significantly related to task completion time.[4] This finding emphasizes the requirement for machines to dialogue about task management and is underscored when comparing the relatively simple CSAR task to many of the domains envisioned for human-machine teams, which frequently are fast paced and high stakes. Articulation work serves as the adaptation to the new demands and negotiation of new task decomposition or new dependencies between teammates. For machines to contribute to articulation work, they must be able to discuss how to change and be able to change. However, articulation work is unsupported because typical dialogue systems have an ontology that is missing task management concepts and have a static task model implemented. Properly supporting articulation has large implications for how dialogue systems, and intelligent systems more generally, must be architected if they are to participate in teamwork. Task models must be modular, such that the task sequence or task decomposition can be adapted to meet the situational demands.

[4]The LIWC measures are not subject to the previous discussion on word count, as they are normalized for word count.

9.5 Future Work

The aim of this work was to identify the general model of common ground by completing the same analysis on several tasks. Coordination common ground models were successful on all four different communication tasks, however, more tasks should be included in future research to further the claims of generality. There was only one asymmetric communication task (the HCRC Map Task) so future work should incorporate asymmetric dialogue settings. Also the majority of tasks had a task completion time metric and the results differed between completion time and accuracy. Future work should explore these differences by additional analysis of tasks with an accuracy measure, or both time and accuracy.

Future work should also situate this research within the broader social sciences work on coordination. Coordination is arguably a stretched term. It has been used in many disciplines and sub-disciplines with different meanings. Here it is used to refer to complementarity (of which there are many types or mechanisms of complementarity). Pickering and Garrod (2004) acknowledge the confusion. Some use coordination to refer to any relationship between interlocutors, of which alignment/similarity is a special type, but Pickering and Garrod restrict coordination in a similar fashion as the current research. Outside of psycholinguistics, for instance in fields of strategic interaction and game theory, coordination is also a common topic of study. Future work should conduct a broad review of coordination in its many forms across social sciences research. It would be valuable to relate the language issues to game theory's signaling and other relevant work.

The results indicating non-stationarity also merit further investigation. First, follow on analyses should investigate if the processes observed are non-stationary and if not, at what time scale (i.e., observation length) they are stationary. The non-stationarity reported here was found for the recurrence quantification analysis processes used in the coordination model. This recurrence plot includes the line of identity, which varies with trial word count and would influence the metrics calculated about lines: average line length (L) and determinism (DET). (Recall that L and DET were the metrics that most strongly indicated

non-stationarity.) In addition, there is likely an interaction between context (e.g., task characteristics) and stationarity. This would be expected particularly if the law of requisite variety accounts for the behaviors observed, as the task increases in complexity so will the communications.

The results clearly supported the coordination model of common ground, and there was partial support for the audience design variant of coordination. Research should further investigate which variant of coordination is present to provide more definite evidence.

9.5.1 Strengthening the Method

Central to the identification of the audience design variant, the statistical mediation analysis related recurrence-based models to Track 2 dialogue. Yet the conclusions from the statistical mediation analysis were limited by the validity of the LIWC-based model of Track 2 dialogue. Future research should develop a better model of Track 2 dialogue, perhaps based on dialogue act annotations, to use for investigating the variant of coordination. Dialogue acts could provide a better model of Track 2 dialogue than the LIWC model. Clark (1996) argued that Track 2 is always occurring in parallel to Track 1, so in some way every utterance has a Track 2 aspect. But there are periods that are dominated by Track 2 because Track 1 is stalled until understanding gets resolved. It is these instances and corresponding dialogue acts that could comprise a more valid model of Track 2 dialogue.

Additionally, the current study showed a preliminary relationship between common ground and articulation work, using only the CSAR task. This task only elicited a sliver of the varied articulation work construct. Future work should explore articulation work more broadly by selecting tasks that require larger and more frequent team articulations. Future work should also get specific about articulation work and create a taxonomy to detail the specific connections to dialogue, as suggested by Rothwell and Shalin (2017).

Another area for research is the recurrence analyses of continuous data, which appeared in the pitch-level and rhythm-level models. Work needed here is refining the method

to improve their construct validity. This work should focus on decreasing their interrelation and providing additional tests of validity using other measures of prosody and rhythmic entrainment (e.g., Levitan and Hirschberg, 2011).

The pitch-level of analysis is also relevant to an issue of Track 2 dialogue when combined with the word-level. For instance, a speaker's utterance followed by a verbatim repeat can serve different Track 2 functions even when it appears to be alignment. With a statement intonation, a verbatim repeat seems like alignment but can also be strong intentional display of understanding. With a question intonation, a verbatim repeat can be either a request for confirmation of what was heard or a request for expansion. The question *"put the water in the bowl?"* can be asking for confirmation that bowl is the correct destination as opposed to other possible destinations or *"put the water in the bowl?"* could be asking for an explanation of the purpose of this command. Linguists have long recognized the interface between prosody and semantics, and the recurrence analysis should be extended to investigate these questions.

The literature separates dyadic communication (only two participants) from multi-party dialogue (three or more). The models of common ground currently are limited to dyads because the recurrence quantification analysis examines two time series. Recently, researchers have extended recurrence quantification analysis to more than two time series (Xu and Yu, 2016). Their approach could be applied to the current models of common ground to investigate multi-party dynamics.

9.5.2 Future Applicability

The support found for coordination and specifically audience design motivate how a dialogue system can model its interlocutors. Historically, symbolic computational models have been along these lines but recently reinforcement learning techniques have proposed new ways for Theory of Mind (Rabinowitz et al., 2018). Also, research in language technology (and human communication for that matter) has underemphasized speech genera-

tion and speech production relative to speech understanding and comprehension. Future research can examine how a model of an interlocutor informs language generation, such as referential expression generation, self-monitoring, and self-repair.

Future work is also merited for the length of installments. Machines need to break up contributions and machines need to accept installments in anticipation of likely input. For machines to break up contributions raises research questions to address: such as when it is necessary to break into installments, and how to break a long contribution into installments.

Finally, research should examine recurrence quantification models of common ground in human interaction with dialogue systems. Demonstrating support for coordination with actual human-computer dialogues would solidify the implications for dialogue system development.

Conclusion

Common ground is a key feature of human dialogue and a necessary capability for machines that use language. The current research examined opposing grounding theories and their implications for the design of dialogue systems. The research simultaneously examined grounding models and task performance in four different tasks at multiple levels of analysis. The findings showed how communication processes contribute to task performance and confirmed the value of recurrence-based models of communication. Findings also clearly supported the coordination model over the alignment model. Further analyses indicated that coordination may result from audience design, the intentional incorporation of the audience's knowledge and perspective. Dialogue system developers should design coordination functions and audience design functions when possible. Systems that rely solely on alignment will have inherent limitations in supporting common ground. Also, the current research emphasized how teams dialogue about articulation work. Articulation work is unsupported by dialogue systems and large-scale architecture changes are required to begin supporting it.

References

Allen, J. F., Schubert, L. K., Ferguson, G., Heeman, P., Hwang, C. H., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M., and Traum, D. R. (1995). The TRAINS project: a case study in building a conversational planning agent.

Anderson, a. H., Badger, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowto, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech*, 34(4), 351–366.

Angus, D., Smith, A., and Wiles, J. (2012a). Conceptual Recurrence Plots : Revealing Patterns in Human Discourse. *IEEE Transactions on Visualization and Computer Graphics*, 18(6), 988–997.

Angus, D., Smith, A., and Wiles, J. (2012b). Human Communication as Coupled Time Series : Quantifying Multi-Participant Recurrence. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6), 1795–1807.

Ashby, W. (1958). Requisite variety and its implications for the control of complex systems. *Cybernetica*, 1, 83–99.

Baker, R. and Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, 43(3), 761–770.

Bangerter, A. and Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive Science*, 27(2), 195–225.

Baron, R. M. and Kenny, D. a. (1986). The Moderator-Mediator Variable Distinction in Social The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182.

Bartel-Radic, A. and Lesca, N. (2011). Do intercultural teams need requisite variety to be effective? *Management International*, 15(3), 89.

Boersma, P. and Weenink, D. (2001). Praat speech processing software. *Institute of Phonetics Sciences of the University of Amsterdam. http://www.praat.org*.

Boghrati, R., Hoover, J., Johnson, K. M., Garten, J., and Dehghani, M. (2017). Conversation level syntax similarity metric. *Behavior Research Methods*.

Bolia, R. S., Nelson, W. T., Ericson, M. a., and Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, 107(2), 1065–1066.

Branigan, H. P., Lickley, R. J., and McKelvie, D. (1999). Non-linguistic influences on rates of disfluency in spontaneous speech. *Proceedings of the 14th International Conference of Phonetic Sciences*, pages 387–390.

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, 75(2), B13–25.

Branigan, H. P., Pickering, M. J., Pearson, J., and McLean, J. F. (2010). Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9), 2355–2368.

Branigan, H. P., Pickering, M. J., Person, J., McLean, J. F., and Nass, C. I. (2003). Syntactic

Alignment Between Computers and People : The Role of Belief about Mental States. *Brain*, 28(9), 186–191.

Brennan, S. E. (1998). The Grounding Problem in Conversations With and Through Computers. In Fussell, S. R. and Kreuz, R. J., editors, *Social and Cognitive Approaches to Interpersonal Communication*, pages 201–225. Lawrence Erlbaum, Hillsdale, NJ.

Brennan, S. E. (2000). Processes that shape conversation and their implications for computational linguistics. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL '00*, pages 1–11.

Brennan, S. E. (2004). How Conversation Is Shaped by Visual and Spoken Evidence. *Approaches to Studying World-Situated Language Use*, pages 95–130.

Brennan, S. E. and Schober, M. F. (2001). How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language*, 44(2), 274–296.

Brennan, S. E., Schuhmann, K. S., and Batres, K. M. (2013). Collaboratively Setting Perspectives and Referring to Locations Across Multiple Contexts. In *Proceedings of the PRE-CogSci 2013 Workshop on the Production of Referring Expressions*, volume 1, pages 1–6, Berlin, Germany.

Brockmann, C., Isard, A., Oberlander, J., and White, M. (2005). Modelling Alignment for Affective Dialogue. In *Proceedings of the Workshop on Adapting the Interaction Style to Affective Factors at the 10th International Conference on User Modeling (UM-05)*, pages 1–5.

Bunt, H., Morante, R., and Keizer, S. (2007). An empirically based computational model of grounding in dialogue. In *Proceedings of the 8th SIGDial Workshop on Discourse and Dialogue*, pages 283–290.

Burtscher, M. J., Wacker, J., Grote, G., and Manser, T. (2010). Managing nonroutine events in anesthesia: The role of adaptive coordination. *Human Factors*, 52(2), 282–294.

Buschmeier, H., Bergmann, K., and Kopp, S. (2009). An Alignment-capable Microplanner for Natural Language Generation. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 82–89.

Butchibabu, A., Sparano-Huiban, C., Sonenberg, L., and Shah, J. (2016). Implicit Coordination Strategies for Effective Team Communication. *Human Factors*, 58(4), 595–610.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-sneddon, G., and Anderson, A. (1996). HCRC Dialogue Structure Coding Manual. Technical report, Human Communication Research Centre, University of Edinborough.

Carletta, J., Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G., and Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Journal Computational Linguistics*, 23(1), 13–31.

Carroll, J. M., Neale, D. C., Isenhour, P. L., Rosson, M. B., and McCrickard, D. S. (2003). Notification and awareness: Synchronizing task-oriented collaborative activity. *International Journal of Human-Computer Studies*, 58, 605–632.

Clark, H. H. (1994). Managing problems in speaking. *Speech Communication*, 15(3-4), 243–250.

Clark, H. H. (1996). *Using Language*. Cambridge University Press, Cambridge, UK.

Clark, H. H. and Brennan, S. E. (1991). Grounding in Communication. In Resnick, L., Levine, J., and Teasley, S., editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, Washington, D.C.

Clark, H. H. and Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language*, 50(1), 62–81.

Clark, H. H. and Marshall, C. R. (1981). Definite reference and mutual knowledge. In Joshi, A. K., Webber, B. L., and Sag, I. A., editors, *Elements of discourse understanding*, pages 10–63. Cambridge University Press, Cambridge, UK.

Clark, H. H. and Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2(1), 19–41.

Clark, H. H. and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, 13(2), 259–294.

Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1–39.

Coco, M. I. and Dale, R. (2014). Cross-recurrence quantification analysis of categorical and continuous time series: an R package. *Frontiers in Psychology*, 5(June), 1–14.

Coco, M. I., Dale, R., and Keller, F. (2017). Performance in a Collaborative Search Task: The Role of Feedback and Alignment. *Topics in Cognitive Science*, 10, 55–79.

Convertino, G., Mentis, H., Rosson, M. B., Slavkovic, A., and Carroll, J. M. (2009). Supporting content and process common ground in computer-supported teamwork. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, number 1, pages 2339–2348.

Convertino, G., Mentis, H. M., Rosson, M. B., Carroll, J. M., Slavkovic, A., and Ganoe, C. H. (2008). Articulating common ground in cooperative work: content and process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1637–1646.

Cooke, N. J. and Gorman, J. C. (2009). Interaction-Based Measures of Cognitive Systems. *Journal of Cognitive Engineering and Decision Making*, 3(1), 27–46.

Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., and Beale, R. (2015). Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human-computer dialogue. *International Journal of Human Computer Studies*, 83, 27–42.

Dale, R., Fusaroli, R., Duran, N. D., and Richardson, D. C. (2014). The Self-Organization of Human Interaction. In *Psychology of Learning and Motivation*, volume 59, pages 43–84.

Dale, R. and Spivey, M. J. (2005). Categorical Recurrence Analysis of Child Language. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, pages 530–535. Lawrence Erlbaum, Mahwah, NJ.

Dale, R. and Spivey, M. J. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, 56(3), 391–430.

DeVault, D. (2008). Contribution tracking: Participating in task-oriented dialogue under uncertainty. *ProQuest Dissertations and Theses*, 3349875, 237.

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., Manrique, E., Rossi, G., and Enfield, N. J. (2015). Universal principles in the repair of communication problems. *PLoS ONE*, 10(9), 1–15.

Duran, N. D., Dale, R., and Kreuz, R. J. (2011). Listeners invest in an assumed other's perspective despite cognitive cost. *Cognition*, 121(1), 22–40.

Dutch Safety Board (2010). Crashed during approach, Boeing 737-800, near Amsterdam Schiphol Airport, 25 February 2009. *Dutch Safety Board Final Report*, page 228.

Fischer, U., McDonnell, L., and Orasanu, J. (2007). Linguistic correlates of team performance: Toward a tool for monitoring team functioning during space missions. *Aviation Space and Environmental Medicine*, 78(5 II).

Fusaroli, R., Raczaszek-Leonardi, J., and Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, 32(1), 147–157.

Fusaroli, R. and Tylén, K. (2016). Investigating Conversational Dynamics: Interactive Alignment, Interpersonal Synergy, and Collective Task Performance. *Cognitive Science*, 40(1), 145–171.

Gabsdil, M. (2003). Clarification in Spoken Dialogue Systems. In *Proceedings of the 2003 AAAI Spring Symposium. Workshop on Natural Language Generation in Spoken and Written Dialogue*, pages 28–35.

Gallagher, S. and Miyahara, K. (2012). Neo-pragmatism and Enactive Intentionality. In Schulkin, J., editor, *Action, Perception and the Brain*, pages 117–146. Palegrave-Macmillan, Basingtoke, UK.

Garland, A. (2015). Ex Machina.

Gergle, D., Kraut, R. E., and Fussell, S. R. (2004). Language Efficiency and Visual Technology: Minimizing Collaborative Effort with Visual Information. *Journal of Language and Social Psychology*, 23(4), 491–517.

Gergle, D., Kraut, R. E., and Fussell, S. R. (2013). Using Visual Information for Grounding and Awareness in Collaborative Tasks. *Human-Computer Interaction*, 28(June 2013), 1–39.

Gervits, F., Eberhard, K., and Scheutz, M. (2016). Team communication as a collaborative Process. *Frontiers in Robotics and AI*, 3(October), 1–14.

Gorman, J. C., Cooke, N. J., Amazeen, P. G., and Fouse, S. (2012). Measuring Patterns in Team Interaction Sequences Using a Discrete Recurrence Approach. *Human Factors*, 54(4), 503–517.

Gorman, J. C., Cooke, N. J., and Kiekel, P. a. (2004). Dynamical Perspectives on Team Cognition. *Cognitive Engineering and Decision Making*, pages 673–677.

Gorman, J. C., Dunbar, T. A., Grimm, D., and Gipson, C. L. (2017). Understanding and modeling teams as dynamical systems. *Frontiers in Psychology*, 8(Jul), 1–18.

Gravano, A. (2009). *Turn-taking and affirmative cue words in task-oriented dialogue*. PhD thesis, Columbia University.

Grimm, D. A., Gorman, J. C., Stevens, R. H., Galloway, T. L., Willemsen-Dunlap, A. M., and Halpin, D. J. (2017). Demonstration of a Method for Real-time Detection of Anomalies in Team Communication. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1), 282–286.

Guastello, S. J. (2017). Nonlinear dynamical systems for theory and research in ergonomics. *Ergonomics*, 60(2), 167–193.

Hampton, A. (2013). *Spatialized Audio and Landmarks in Team Navigation*. PhD thesis, Wright State University.

Hampton, A., Shalin, V. L., Robinson, E., Simpson, B. D., Finomore, V., Cowgill, J., Moore, T., Rapoch, T., and Gilkey, R. (2012). The Impact of Spatialized Communications on Team Navigation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1), 463–467.

Hawkins, R. X. D. and Goodman, N. D. (2016). Conversational expectations account for apparent limits on theory of mind use. In *Proceedings of the 38th Conference of the Cognitive Science Society*, pages 1889–1894.

Healey, P. G. T., Purver, M., and Howes, C. (2014). Divergence in dialogue. *PloS one*, 9(2), e98598.

Heath, C. and Luff, P. (1992). Collaboration and Control: Crisis Management and Multimedia Technology in London Underground Line Control Rooms. *Computer Supported Cooperative Work*, 1(1990), 69–94.

Heeman, P. A. and Allen, J. F. (1998). Speech Repairs, Intonational Boundaries and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialog. *Computational Linguistics*, 25(4), 527–571.

Hirst, G., McRoy, S., Heeman, P., Edmonds, P., and Horton, D. (1994). Repairing conversational misunderstandings and non-understandings. *Speech Communication*, 15(3-4), 213–229.

Hlavac, M. (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1.

Honnibal, M. and Johnson, M. (2014). Joint Incremental Disfluency Detection and Dependency Parsing. *Transactions of the Association for Computational Linguistics*, 2, 131–142.

Horton, W. S. and Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, 96(2), 127–142.

Horton, W. S. and Keysar, B. (1996). When do speakers take into common ground? *Cognition*, 56(1), 91–117.

Iyer, N., Thompson, E., Stillwagon, K., Ennis, Z., Willis, A., and Simpson, B. (2016). Adaptive speech modifications and its effect on communication effectiveness in complex acoustic environments. *The Journal of the Acoustical Society of America*, 140(4), 3437.

Janarthanam, S. and Lemon, O. (2009). A wizard-of-Oz environment to study referring expression generation in a situated spoken dialogue task. In *Proceedings of the 12th European Workshop on Natural Language Generation, ENLG 2009*, number March, pages 94–97.

Johnson, M. and Charniak, E. (2004). A TAG-based noisy channel model of speech repairs. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, pages 33–es.

Jonze, S. (2013). Her.

Kantz, H. and Schreiber, T. (2004). *Nonlinear Time Series Analysis*.

Keysar, B., Barr, D. J., Balin, J. A., and Brauner, J. S. (2000). Taking Perspective in Conversation: The Role of Mutual Knowledge in Comprehension. *Psychological Science*, 11(1), 32–38.

Keysar, B. and Horton, W. S. (1998). Speaking with common ground: from principles to processes in pragmatics: a reply to Polichak and Gerrig. *Cognition*, 66(2), 191–198.

Khawaja, M. A., Chen, F., and Marcus, N. (2012). Analysis of collaborative communication for linguistic cues of cognitive load. *Human Factors*, 54(4), 518–529.

Khawaja, M. A., Chen, F., and Marcus, N. (2014). Measuring Cognitive Load Using Linguistic Features: Implications for Usability Evaluation and Adaptive Interaction Design. *International Journal of Human-Computer Interaction*, 30(5), 343–368.

Kiekel, P. a., Cooke, N. J., Foltz, P. W., Gorman, J. C., and Martin, M. J. (2002). Some Promising Results of Communication-Based Automatic Measures of Team Cognition. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 46(3), 298–302.

Koulouri, T. and Lauria, S. (2009). Exploring miscommunication and collaborative behaviour in human-robot interaction. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 111–119. Association for Computational Linguistics.

Koulouri, T., Lauria, S., and Macredie, R. D. (2015). Do (and Say) as I Say: Linguistic Adaptation in HumanComputer Dialogs. *Human-Computer Interaction*, 24(January), 59–95.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1), 41–104.

Levitan, R. and Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (August), 3081–3084.

Lickley, R. (2001). Dialogue moves and disfluency rates. *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech (DiSS'01)*, pages 1–4.

Louwerse, M. M., Dale, R., Bard, E. G., and Jeuniax, P. (2012). Behavior Matching in Multimodal Communication Is Synchronized. *Cognitive Science*, 36(8), 1404–1426.

Malone, T. W. and Crowston, K. (1990). What is coordination theory and how can it help design cooperative work systems? *Proceedings of the 1990 ACM Conference on Computer-Supported Cooperative Work - CSCW '90*, (April), 357–370.

Marwan, N., Carmen Romano, M., Thiel, M., and Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6), 237–329.

Mills, G. J. (2014). Dialogue in joint activity: Complementarity, convergence and conventionalization. *New Ideas in Psychology*, 32(1), 158–173.

Mooney, R. (2008). Learning to Connect Language and Perception. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, (July), 1598–1601.

Moscoso del Prado Martín, F. and Du Bois, J. (2015). Syntactic Alignment is an Index of Affective Alignment : An Information-Theoretical Study of Natural Dialogue. In *Cognitive Science Society Annual Meeting*, pages 1655–1660.

National Transporation Safety Board (2014). Crash of Asiana Flight 214 Accident Report Summary.

Olmstead, K. (2017). A third of Americans live in a household with three or more smartphones.

Orsucci, F., Petrosino, R., Paoloni, G., Canestri, L., Conte, E., Reda, M., and Fulcheri, M. (2013). Prosody and synchronization in cognitive neuroscience. *EPJ Nonlinear Biomedical Physics*, 1(1), 1:6.

Oser, R. L., Prince, C., Morgan Jr, B. B., and Simpson, S. S. (1991). An analysis of aircrew communication patterns and content. Technical report, Naval Traning Systems Center.

Oviatt, S. (1995). Predicting spoken disfluencies during humancomputer interaction. *Computer Speech & Language*, 9(1), 19–35.

Passonneau, R. J., Epstein, S. L., and Ligorio, T. (2012). Naturalistic Dialogue Management for Noisy Speech Recognition. *Journal of Selected Topics in Signal Processing Special Issue*, 6(8), 928–942.

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The Development and Psychometric Properties of LIWC2015. Technical report, University of Texas at Austin, Austin, TX.

Pew Research Center (2014). The Internet of Things Will Thrive by 2025. Technical Report May.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27, 169–190; discussion 190–226.

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M. A., and Botvinick, M. (2018). Machine Theory of Mind. *arXiv preprint*.

Raczaszek-Leonardi, J. (2016). Multiple Systems and Multiple Time Scales of Language Dynamics : coping with complexity. *Cybernetics and Human Knowing*, 21(October), 37–52.

Raczaszek-Leonardi, J., Debska, A., and Sochanowicz, A. (2014). Pooling the ground: Understanding and coordination in collective sense making. *Frontiers in Psychology*, 5(OCT), 1–13.

Reitter, D. and Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76, 29–46.

Rieke, C., Andrzejak, R. G., Mormann, F., and Lehnertz, K. (2004). Improved statistical test for nonstationarity using recurrence time statistics. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 69(4), 9.

Romigh, G. D., Rothwell, C. D., Greenwell, B., and Newman, M. (2016). Modeling uncertainty in spontaneous speech: Lexical and acoustic features. In *Poster session presented at the 172nd meeting of the Acoustical Society of America.*, Honolulu, HI.

Roque, A. and Traum, D. (2008). Degrees of grounding based on evidence of understanding. *Proceedings of the SIGDIAL 2008 Conference: The 9th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 54–63.

Rothwell, C. D. and Shalin, V. L. (2017). Human-Machine Articulation Work: Functional Dependency Dialogue for Human-Machine Teaming. In *Proceedings of the The 19th International Symposium of Aviation Psychology*, pages 1–6.

Salas, E., DiazGranados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., and

Halpin, S. M. (2008). Does team training improve team performance? A meta-analysis. *Human Factors*, 50(6), 903–933.

Schaller, S. (2012). *A Man Without Words*. University of California Press, Oakland, CA.

Schegloff, E. A. and Sacks, H. (2006). Opening Up Closings. In Jaworksi, A. and Coupland, N., editors, *The Discourse Reader*, pages 262–271. Routledge, New York, NY, 2nd edition.

Schmidt, K. and Bannon, L. (1992). Taking CSCW Seriously: Supporting Articulation Work. *Computer Supported Cooperative Work*, 1(1), 7–40.

Schober, M. F. and Brennan, S. E. (2003). Processes of Interactive Spoken Discourse: The Role of the Partner. In Graesser, A. C., Gernsbacher, M. A., and Goldman, S. R., editors, *Handbook of Discourse Processes*, pages 123–164. Routledge.

Searle, J. R. (1990). Is the Brain's Mind a Computer Program? *Scientific American*, 262, 26–31.

Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California at Berkeley.

Shriberg, E. (2005). Spontaneous speech: How people really talk and why engineers should care. *Interspeech 2005, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 1781–1784.

Silva, S. S. and Hansman, R. J. (2015). Divergence Between Flight Crew Mental Model and Aircraft System State in Auto-Throttle Mode Confusion Accident and Incident Cases. *Journal of Cognitive Engineering and Decision Making*, 9(4), 312–328.

Skantze, G. (2005). Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(3), 325–341.

Strauss, A. (1985). Work and the Division of Labor. *The Sociological Quarterly*, 26(1), 1–19.

Svensson, J. and Andersson, J. (2006). Speech acts, communication problems, and fighter pilot team performance. *Ergonomics*, 49(12-13), 1226–1237.

Taylor, A., Marcus, M., and Santorini, B. (2003). The Penn Treebank: an Overview. In Abeille, A., editor, *Treebanks: Building and Using Parsed Corpora*, pages 5–22.

Tenbrink, T., Andonova, E., and Coventry, K. (2008). Negotiating spatial relationships in dialogue: The role of the addressee. In *Proceedings of LONDIAL - The 12th SEMDIAL workshop*, pages 193–200.

Thomason, J., Sinapov, J., Svetlik, M., Stone, P., and Mooney, R. J. (2016). Learning multi-modal grounded linguistic semantics by playing "I spy". *IJCAI International Joint Conference on Artificial Intelligence*, 2016-Janua, 3477–3483.

Tomko, S. L. (2006). *Improving User Interaction with Spoken Dialog Systems via Shaping*. PhD thesis, Carnegie Mellon University.

Traum, D. R. (1994). *A computational theory of grounding in natural language conversation*. PhD thesis, University of Rochester.

Traum, D. R. and Dillenbourg, P. (1996). Miscommunication in Multi-modal Collaboration. In *AAAI Workshop on Detecting, Repairing, And Preventing Human–Machine Miscommunication*, pages 37–46.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., and Bradlow, A. R. (2010). The Wildcat Corpus of native- and foreign-accented English: communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*, 53(Pt 4), 510–40.

Varges, S. (2006). Overgeneration and ranking for spoken dialogue systems. In *INLG 2006*, pages 3–5.

Visser, T. (2011). Toward a Model for Incremental Grounding in Dialogue Systems. In Böck, R., Bonin, F., Campbell, N., Edlund, J., de Kok, I. A., Poppe, R. W., and Traum, D. R., editors, *Joint Proceedings of the Intelligent Virtual Agents 2012 Workshops*, pages 101–108. Otto von Guericke University.

VoiceLabs.co (2017). The 2017 Voice Report: Executive Summary. Technical report.

Ward, N. G. and Devault, D. (2015). Ten Challenges in Highly-Interactive Dialog Systems. In *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*, pages 104–107.

Weatherholtz, K., Campbell-Kibler, K., and Jaeger, T. F. (2014). Socially-mediated syntactic alignment. *Language Variation and Change*, 26, 387–420.

Weber, R. a. and Camerer, C. F. (2003). Cultural Conflict and Merger Failure: An Experimental Approach. *Management Science*, 49(4), 400–415.

Xu, T. and Yu, C. (2016). Quantifying Joint Activities using Cross-Recurrence Block Representation. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, pages 1997–2002.

Yee, D. J., Wiggins, M. W., and Searle, B. J. (2017). The Role of Social Cue Utilization and Closing-the- Loop Communication in the Performance of Ad Hoc Dyads. *Human Factors*, 59(6), 1009–1021.

Appendix A: Recurrence Metric Chance Analyses

Tables A.1 and A.2 show detailed results of the chance analyses performed for each task, for each level of analysis, for each model, and for each recurrence metric.

Table A.1: Results of chance analyses using shuffled controls, showing p-values from two-sided paired t-tests. All Map task tests had df = 127. All Uncertainty tests had df = 39.

Map Task								
	Prosody Level				*Rhythm Level*			
	RR	*DET*	*L*	*ENTR*	*RR*	*DET*	*L*	*ENTR*
Alignment	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Coordination	0.61	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Baseline	$< .05$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
	Morpheme Level				*Word Level*			
Alignment	$< .001$	$< .001$	$< .001$	$< .001$	N/A	$< .001$	$< .001$	$< .001$
Coordination	$< .001$	$< .001$	$< .001$	$< .001$	N/A	$< .001$	$< .001$	$< .001$
Baseline	$< .001$	$< .001$	$< .001$	$< .001$	N/A	$< .001$	$< .001$	$< .001$
	Syntax Level							
Alignment	N/A	$< .001$	$< .001$	$< .001$				
Coordination	N/A	$< .001$	$< .001$	$< .001$				
Baseline	N/A	$< .001$	$< .001$	$< .001$				
Uncertainty Task								
	Prosody Level				*Rhythm Level*			
	RR	*DET*	*L*	*ENTR*	*RR*	*DET*	*L*	*ENTR*
Alignment	–	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Coordination	–	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Baseline	–	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
	Morpheme Level				*Word Level*			
Alignment	$< .001$	$< .001$	$< .001$	$< .001$	N/A	$< .001$	$< .001$	$< .001$
Coordination	$< .001$	$< .001$	$< .001$	$< .001$	N/A	$< .001$	$< .001$	$< .001$
Baseline	$< .001$	$< .001$	$< .001$	$< .001$	N/A	$< .001$	$< .001$	$< .001$
	Syntax Level							
Alignment	N/A	$< .001$	$< .001$	$< .001$				
Coordination	N/A	$< .001$	$< .001$	$< .001$				
Baseline	N/A	$< .001$	$< .001$	$< .001$				

Table A.2: Results of chance analyses using shuffled controls, showing *p*-values from two-sided paired t-tests. The non-significant tests are shown in bold. All Diapix task tests had *df* = 47, except the word-level alignment metrics DET and ENTR had df = 39 because no lines existed in 8 of the shuffled recurrence plots. All CSAR tests had *df* = 119.

Diapix Task								
	Prosody Level				*Rhythm Level*			
	RR	*DET*	*L*	*ENTR*	*RR*	*DET*	*L*	*ENTR*
Alignment	0.32	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Coordination	0.68	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Baseline	0.12	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
	Morpheme Level				*Word Level*			
Alignment	$< .001$	**0.12**	$< .001$	$< .001$	N/A	$< .001$	$< .001$	$< .001$
Coordination	$< .001$	$< .001$	$< .001$	$< .001$	N/A	$< .001$	$< .001$	$< .001$
Baseline	$< .001$	$< .001$	$< .001$	**0.28**	N/A	$< .001$	$< .001$	$< .001$
	Syntax Level							
Alignment	N/A	$< .001$	$< .001$	$< .001$				
Coordination	N/A	$< .001$	$< .001$	$< .001$				
Baseline	N/A	$< .001$	$< .01$	$< .001$				
CSAR Task								
	Prosody Level				*Rhythm Level*			
	RR	*DET*	*L*	*ENTR*	*RR*	*DET*	*L*	*ENTR*
Alignment	–	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Coordination	–	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
Baseline	–	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
	Morpheme Level				*Word Level*			
Alignment	$< .001$	**0.18**	**0.35**	**0.26**	N/A	$< .001$	$< .001$	$< .001$
Coordination	$< .001$	$< .001$	$< .001$	$< .001$	N/A	$< .001$	$< .001$	$< .001$
Baseline	$< .001$	$< .001$	$< .001$	$< .001$	N/A	$< .001$	$< .001$	$< .001$
	Syntax Level							
Alignment	N/A	$< .001$	$< .001$	$< .001$				
Coordination	N/A	$< .001$	$< .001$	$< .001$				
Baseline	N/A	$< .001$	$< .001$	$< .001$				

Appendix B: Complete Recurrence Models

Table B.1: Uncertainty task, Alignment Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	23,114,903.000	−58.522	17,744.390**	166.341	−123.545
	(56,749,640.000)	(667.813)	(8,053.951)	(1,253.742)	(456.987)
DET	8,098.399	928.535	−32.526***	−12.984	14.341
	(11,071.290)	(552.319)	(11.832)	(17.020)	(17.688)
L	−609.892*	−3.131	−3,014.456*	−7,824.701**	−23,865.690*
	(311.483)	(5.697)	(1,492.349)	(3,313.950)	(13,660.110)
ENTR	2,927.824	−463.384***	4,085.366**	5,021.582**	15,450.430*
	(2,046.387)	(130.609)	(1,917.471)	(1,885.303)	(8,881.913)
Constant	−935,065.400	1,415.037***	5,455.369**	15,428.550**	44,597.550*
	(2,271,004.000)	(159.845)	(2,094.374)	(6,480.018)	(25,494.770)
Observations	36	38	38	38	38
Adjusted R^2	0.038	0.678	0.228	0.099	0.025
Residual Std. Error	239.093 (df = 31)	159.926 (df = 33)	247.689 (df = 33)	267.627 (df = 33)	278.364 (df = 33)
F Statistic	1.349 (df = 4; 31)	20.506*** (df = 4; 33)	3.738** (df = 4; 33)	2.019 (df = 4; 33)	1.242 (df = 4; 33)
Note:					*p<0.1; **p<0.05; ***p<0.01

Table B.2: Uncertainty task, Coordination Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	525,033.000	548.192	−1,129.366	−783.636***	−81.842
	(416,228.900)	(569.074)	(761.541)	(178.207)	(105.078)
DET	6,284.413*	723.057	−26.140***	−2.331	−29.034*
	(3,519.224)	(482.680)	(6.776)	(9.063)	(16.819)
L	−252.486	7.292	−701.770***	−516.260***	−5,373.647***
	(884.329)	(17.987)	(94.459)	(84.316)	(634.747)
ENTR	−694.414	−664.545***	1,842.669***	614.001**	4,102.829***
	(3,227.918)	(184.232)	(342.911)	(242.993)	(909.017)
Constant	−23,138.130	1,426.580***	2,806.875***	3,135.208***	12,248.780***
	(15,911.380)	(152.322)	(360.686)	(298.430)	(1,400.551)
Observations	36	38	38	38	38
Adjusted R^2	0.258	0.704	0.763	0.745	0.675
Residual Std. Error	210.067 (df = 31)	153.306 (df = 33)	137.413 (df = 33)	142.446 (df = 33)	160.635 (df = 33)
F Statistic	4.037*** (df = 4; 31)	23.044*** (df = 4; 33)	30.701*** (df = 4; 33)	27.997*** (df = 4; 33)	20.253*** (df = 4; 33)
Note:					*p<0.1; **p<0.05; ***p<0.01

Table B.3: Uncertainty task, Baseline Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	−50,425.850	−237.279	−532.128	−274.668**	32.708
	(76,085.590)	(761.122)	(795.925)	(124.163)	(57.802)
DET	7,557.256**	783.785	−28.099***	−18.287**	−29.878**
	(2,960.437)	(663.662)	(8.545)	(6.823)	(11.794)
L	−1,276.300*	−0.215	−97.529**	−65.230*	−2,227.327***
	(668.414)	(4.618)	(39.093)	(38.432)	(351.530)
ENTR	2,445.137	−402.895***	409.512	153.157	1,672.659***
	(2,527.277)	(136.321)	(274.670)	(199.595)	(525.312)
Constant	−2,986.188	1,430.843***	2,665.530***	1,873.690***	5,777.342***
	(2,685.399)	(193.275)	(321.081)	(226.196)	(720.622)
Observations	38	38	38	38	38
Adjusted R^2	0.349	0.530	0.650	0.515	0.563
Residual Std. Error (df = 33)	227.499	193.406	166.729	196.351	186.491
F Statistic (df = 4; 33)	5.961***	11.412***	18.208***	10.827***	12.897***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table B.4: Diapix task, Alignment Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	36,339.080	−470.245**	−443.245	96.128	7.005
	(36,467.780)	(208.675)	(1,040.997)	(138.687)	(24.408)
DET	415.195	−5,250.045	–	−2.133	1.258
	(1,010.247)	(3,624.536)		(1.683)	(2.988)
L	4.461	−0.951	11.806	671.324**	2,080.083**
	(26.720)	(4.279)	(230.451)	(299.601)	(861.521)
ENTR	−6.118	171.572*	35.754	−442.103**	−1,408.644**
	(218.085)	(91.950)	(255.523)	(194.652)	(528.579)
Constant	−1,687.683	5,024.918	126.853	−1,080.019*	−3,685.675**
	(1,696.637)	(3,411.300)	(346.283)	(579.667)	(1,626.173)
Observations	40	40	40	40	40
Adjusted R^2	−0.032	0.233	−0.061	0.066	0.096
Residual Std. Error	51.935 (df = 35)	44.776 (df = 35)	52.643 (df = 36)	49.406 (df = 35)	48.612 (df = 35)
F Statistic	0.696 (df = 4; 35)	3.957*** (df = 4; 35)	0.258 (df = 3; 36)	1.687 (df = 4; 35)	2.031 (df = 4; 35)
Note:	Predictors that were not greater than chance are removed, denoted by –. *p<0.1; **p<0.05; ***p<0.01				

Table B.5: Diapix task, Coordination Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	−143,471.400**	−67.488	46.652	7.613	−9.823
	(53,336.880)	(476.243)	(113.104)	(25.627)	(6.563)
DET	246.160	1,301.668	−0.973	−2.813**	4.412
	(351.828)	(1,268.770)	(0.960)	(1.066)	(2.806)
L	200.230	26.783	−70.585***	−16.279*	−415.428***
	(246.911)	(17.893)	(18.055)	(8.803)	(96.711)
ENTR	−496.798	−258.660	118.746**	20.956	71.037
	(602.374)	(177.600)	(52.836)	(41.123)	(127.280)
Constant	5,827.259***	−561.674	338.006***	295.623***	1,088.558***
	(2,147.729)	(859.783)	(68.742)	(53.678)	(205.824)
Observations	46	46	46	46	46
Adjusted R^2	0.141	0.116	0.285	0.227	0.273
Residual Std. Error (df = 41)	37.268	37.786	33.988	35.335	34.274
F Statistic (df = 4; 41)	2.839**	2.483*	5.488***	4.311***	5.226***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table B.6: Diapix task, Baseline Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	3,130.400	−124.338	−48.895	−33.818**	−14.057***
	(25,810.810)	(205.549)	(155.613)	(14.304)	(5.017)
DET	−114.062	2,678.014	2.552	3.051**	7.166***
	(573.505)	(4,525.886)	(1.773)	(1.221)	(2.471)
L	469.656	0.012	−3.672	−5.000	28.491
	(302.221)	(0.927)	(3.363)	(4.557)	(49.032)
ENTR	−905.450	3.559	–	−29.415	−167.945
	(764.080)	(40.455)		(35.914)	(100.236)
Constant	−8.043	−2,404.264	46.128	189.283***	115.352
	(973.737)	(4,305.624)	(85.958)	(39.811)	(132.531)
Observations	40	40	40	40	40
Adjusted R^2	0.079	−0.086	−0.016	0.078	0.157
Residual Std. Error	49.046 (df = 35)	53.264 (df = 35)	51.536 (df = 36)	49.080 (df = 35)	46.939 (df = 35)
F Statistic	1.841 (df = 4; 35)	0.230 (df = 4; 35)	0.790 (df = 3; 36)	1.826 (df = 4; 35)	2.813** (df = 4; 35)
Note:	Predictors that were not greater than chance are removed, denoted by –. *p<0.1; **p<0.05; ***p<0.01				

Table B.7: Map task Path Deviation, Alignment Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	−1,705.417**	−171.529*	637.234	8.550	8.315
	(765.342)	(100.314)	(410.960)	(45.962)	(14.062)
DET	−98.646	1,174.954*	−0.383	1.523	−1.543
	(89.846)	(654.146)	(0.731)	(1.045)	(1.082)
L	0.184	0.003	164.618**	213.792	158.936
	(3.180)	(2.550)	(79.366)	(180.662)	(242.760)
ENTR	10.650	38.289	−165.596	−138.722	−104.005
	(21.695)	(37.566)	(108.327)	(119.005)	(161.647)
Constant	193.712***	−1,063.499**	−195.296*	−357.773	−201.125
	(71.839)	(525.559)	(104.569)	(344.024)	(457.949)
Observations	124	124	124	123	124
Adjusted R^2	0.025	0.087	0.031	−0.004	0.005
Residual Std. Error	42.600 (df = 119)	41.223 (df = 119)	42.466 (df = 119)	43.227 (df = 118)	43.039 (df = 119)
F Statistic	1.793 (df = 4; 119)	3.936*** (df = 4; 119)	1.992 (df = 4; 119)	0.866 (df = 4; 118)	1.153 (df = 4; 119)
Note:					*p<0.1; **p<0.05; ***p<0.01

Table B.8: Map task Path Deviation, Coordination Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	−17,247.400	−362.246**	29.609	−13.867	−13.593***
	(20,549.720)	(158.296)	(73.926)	(12.581)	(5.034)
DET	14.435	−130.421	0.720	2.851***	0.611
	(120.943)	(310.997)	(0.834)	(0.931)	(1.651)
L	−57.865	8.160	18.266	9.317	472.709***
	(83.963)	(8.632)	(21.584)	(20.647)	(116.129)
ENTR	36.102	105.620	−11.279	−46.791	−373.061***
	(153.172)	(75.892)	(74.969)	(41.028)	(119.462)
Constant	862.713	73.215	−45.727	32.255	−675.831***
	(866.862)	(129.682)	(65.674)	(54.255)	(203.899)
Observations	124	124	124	124	124
Adjusted R^2	−0.004	0.065	0.022	0.063	0.128
Residual Std. Error (df = 119)	43.241	41.729	42.676	41.766	40.284
F Statistic (df = 4; 119)	0.865	3.124**	1.680	3.065**	5.523***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table B.9: Map task Path Deviation, Baseline Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	−13,091.290	−241.581**	85.276	−2.947	−2.940
	(9,357.127)	(110.967)	(69.563)	(2.729)	(1.935)
DET	−6.799	2,137.432**	0.587	1.296*	0.024
	(120.822)	(978.414)	(0.697)	(0.716)	(1.253)
L	−116.892	0.062	−0.469	3.297	100.717***
	(78.532)	(1.122)	(0.592)	(3.449)	(37.307)
ENTR	154.449	−8.124	−6.187	10.203	−54.779
	(118.160)	(24.326)	(25.166)	(27.065)	(48.118)
Constant	760.317*	−1,823.581**	14.103	17.841	−113.150
	(400.709)	(850.847)	(50.955)	(26.558)	(75.167)
Observations	124	124	124	124	124
Adjusted R^2	0.0002	0.031	0.005	0.012	0.030
Residual Std. Error (df = 119)	43.141	42.468	43.042	42.881	42.499
F Statistic (df = 4; 119)	1.006	1.989	1.148	1.381	1.943
Note:					*p<0.1; **p<0.05; ***p<0.01

Table B.10: Map task Completion Time, Alignment Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	−443.550	−56.052	14.464	−91.211	−7.135
	(2,243.575)	(320.879)	(1,195.711)	(131.463)	(42.373)
DET	−87.904	1,228.626	3.372	−7.172**	2.510
	(266.772)	(2,100.551)	(2.134)	(3.023)	(3.290)
L	−9.203	3.029	−964.047***	−2,061.141***	−2,317.459***
	(9.768)	(8.135)	(233.683)	(522.417)	(724.153)
ENTR	150.799**	−258.895**	1,239.540***	1,400.325***	1,453.748***
	(63.746)	(120.783)	(317.900)	(341.931)	(482.290)
Constant	226.366	−94.650	1,406.652***	4,322.534***	4,654.195***
	(210.830)	(1,687.632)	(309.909)	(995.575)	(1,364.926)
Observations	122	122	122	121	122
Adjusted R^2	0.088	0.031	0.121	0.115	0.063
Residual Std. Error	127.684 (df = 117)	131.594 (df = 117)	125.331 (df = 117)	125.385 (df = 116)	129.450 (df = 117)
F Statistic	3.916*** (df = 4; 117)	1.974 (df = 4; 117)	5.173*** (df = 4; 117)	4.888*** (df = 4; 116)	3.017** (df = 4; 117)
Note:					*p<0.1; **p<0.05; ***p<0.01

Table B.11: Map task Completion Time, Coordination Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	388,072.600***	791.448	−915.609***	−155.293***	−39.456***
	(53,070.620)	(493.632)	(164.242)	(23.965)	(11.511)
DET	−411.431	767.254	4.362**	−5.786***	−10.425***
	(313.676)	(996.188)	(1.896)	(1.817)	(3.805)
L	429.948**	−19.500	−284.610***	−345.281***	−2,343.157***
	(213.816)	(26.940)	(47.568)	(39.667)	(271.587)
ENTR	−327.361	−344.389	676.558***	311.845***	2,072.866***
	(388.231)	(242.111)	(163.661)	(77.945)	(274.452)
Constant	−15,656.350***	180.641	462.034***	1,624.115***	5,235.056***
	(2,240.868)	(416.676)	(146.243)	(108.319)	(479.019)
Observations	122	122	122	122	122
Adjusted R^2	0.322	0.037	0.492	0.626	0.520
Residual Std. Error (df = 117)	110.084	131.214	95.273	81.742	92.633
F Statistic (df = 4; 117)	15.369***	2.155*	30.320***	51.673***	33.764***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table B.12: Map task Completion Time, Baseline Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	91,745.370***	−205.837	−967.712***	−7.090	2.228
	(27,926.170)	(313.746)	(173.733)	(6.770)	(4.796)
DET	−359.134	−4,583.718	−1.594	−8.858***	−8.718***
	(363.512)	(2,800.438)	(1.771)	(1.836)	(3.095)
L	88.933	5.244	1.855	−28.708***	−533.360***
	(235.880)	(3.237)	(1.533)	(8.531)	(94.384)
ENTR	120.512	−97.908	110.934*	−1.123	399.235***
	(352.470)	(70.011)	(64.860)	(67.950)	(118.884)
Constant	−3,390.187***	5,198.222**	620.124***	787.622***	1,756.467***
	(1,199.077)	(2,438.510)	(133.029)	(68.567)	(193.854)
Observations	122	122	122	122	122
R^2	0.124	0.205	0.337	0.387	0.393
Adjusted R^2	0.094	0.177	0.314	0.366	0.372
Residual Std. Error (df = 117)	127.283	121.254	110.703	106.425	105.933
F Statistic (df = 4; 117)	4.125***	7.527***	14.871***	18.490***	18.934***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table B.13: CSAR task, Alignment Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	3,449,879.000*	183.192	426.891	−144.425**	−25.960
	(1,772,836.000)	(155.462)	(569.020)	(65.424)	(25.638)
DET	−560.056	−4,070.501	–	−3.085***	1.123
	(375.814)	(2,469.823)		(0.809)	(1.096)
L	−6.054	−0.017	–	−133.800***	−1,514.674***
	(9.750)	(0.399)		(38.360)	(393.816)
ENTR	74.132	−24.983	–	147.506***	975.112***
	(71.128)	(39.083)		(25.551)	(242.358)
Constant	−137,385.800*	4,169.763*	173.693***	510.181***	3,023.214***
	(70,836.850)	(2,339.369)	(18.348)	(81.652)	(760.033)
Observations	106	106	106	104	106
R^2	0.052	0.073	0.005	0.312	0.185
Adjusted R^2	0.014	0.036	−0.004	0.284	0.152
Residual Std. Error	52.671 (df = 101)	52.092 (df = 101)	53.166 (df = 104)	44.208 (df = 99)	48.845 (df = 101)
F Statistic	1.385 (df = 4; 101)	1.980 (df = 4; 101)	0.563 (df = 1; 104)	11.203*** (df = 4; 99)	5.721*** (df = 4; 101)
Note:	Predictors that were not greater than chance are removed, denoted by –. *p<0.1; **p<0.05; ***p<0.01				

Table B.14: CSAR task, Coordination Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	177,967.900***	191.474	−91.788	−38.953***	−5.267
	(25,910.330)	(163.931)	(70.046)	(13.461)	(8.127)
DET	−235.798	−1,269.219*	−2.828***	−2.661***	0.785
	(173.689)	(758.278)	(1.037)	(0.511)	(1.337)
L	−15.034	−0.688	−13.592***	−0.373	−151.875***
	(15.529)	(1.057)	(3.806)	(0.476)	(18.407)
ENTR	99.195	−9.874	32.604	69.624***	84.706*
	(70.501)	(51.026)	(26.558)	(20.996)	(48.879)
Constant	−6,766.709***	1,357.164**	451.242***	308.024***	551.007***
	(998.283)	(630.889)	(70.609)	(21.138)	(68.878)
Observations	106	106	106	106	106
Adjusted R^2	0.318	0.044	0.393	0.451	0.389
Residual Std. Error (df = 101)	43.817	51.888	41.322	39.320	41.485
F Statistic (df = 4; 101)	13.237***	2.195*	18.025***	22.544***	17.685***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table B.15: CSAR task, Baseline Models.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	55,963.100***	86.794	−40.934	−16.184**	−5.198
	(13,477.890)	(115.308)	(30.461)	(6.213)	(6.609)
DET	−79.314	−4,399.356**	−2.945***	−1.068**	−0.715
	(141.022)	(2,058.079)	(0.781)	(0.483)	(0.991)
L	−5.765	0.070	0.112	−0.484	−7.613**
	(12.277)	(0.170)	(0.293)	(0.465)	(3.408)
ENTR	34.885	−4.295	20.040	20.693	−43.773
	(55.242)	(24.596)	(20.280)	(17.392)	(31.585)
Constant	−1,973.490***	4,494.682**	421.229***	275.650***	314.923***
	(499.890)	(1,952.613)	(60.870)	(20.571)	(51.565)
Observations	106	106	106	105	106
Adjusted R^2	0.141	0.063	0.304	0.250	0.079
Residual Std. Error	49.181 (df = 101)	51.350 (df = 101)	44.271 (df = 101)	45.976 (df = 100)	50.910 (df = 101)
F Statistic	5.299*** (df = 4; 101)	2.773** (df = 4; 101)	12.451*** (df = 4; 101)	9.650*** (df = 4; 100)	3.260** (df = 4; 101)
Note:					*p<0.1; **p<0.05; ***p<0.01

Table B.16: Uncertainty task First Submission Accuracy, Alignment Models.

Level:	Morpheme	Word	Syntax
RR	1.276*	1.979	0.0005
	(0.687)	(6.080)	(0.239)
DET	−0.005	0.016**	0.019
	(0.012)	(0.008)	(0.020)
L	−2.481	−0.403	−10.432
	(3.372)	(1.102)	(8.068)
ENTR	1.291	0.141	6.154
	(1.990)	(1.443)	(5.306)
Constant	5.040	0.455	19.691
	(6.461)	(1.436)	(15.190)
Observations	37	37	37
Adjusted R^2	0.052	0.138	0.141
Residual Std. Error (df = 32)	0.205	0.195	0.195
F Statistic (df = 4; 32)	1.492	2.445*	2.478*
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table B.17: Uncertainty task First Submission Accuracy, Coordination Models.

Level:	Morpheme	Word	Syntax
RR	0.136	−0.751	0.009
	(0.191)	(0.992)	(0.100)
DET	−0.005	0.013	−0.020
	(0.012)	(0.008)	(0.020)
L	−0.147	−0.266***	−1.807***
	(0.095)	(0.087)	(0.586)
ENTR	0.355	0.529	2.510**
	(0.316)	(0.392)	(1.035)
Constant	0.725**	0.126	3.774**
	(0.343)	(0.436)	(1.411)
Observations	37	37	37
Adjusted R^2	0.183	0.181	0.227
Residual Std. Error (df = 32)	0.190	0.190	0.185
F Statistic (df = 4; 32)	3.019**	2.992**	3.637**
Note:			*p<0.1; **p<0.05; ***p<0.01

Table B.18: Uncertainty task First Submission Accuracy, Baseline Models.

Level:	Morpheme	Word	Syntax
RR	0.083	−1.254	0.064
	(0.117)	(0.856)	(0.092)
DET	−0.002	0.021***	0.003
	(0.007)	(0.007)	(0.016)
L	−0.067**	−0.059***	−0.218
	(0.029)	(0.017)	(0.136)
ENTR	−0.016	−0.089	0.040
	(0.209)	(0.192)	(0.595)
Constant	0.771***	−0.118	0.481
	(0.196)	(0.350)	(0.658)
Observations	37	37	37
Adjusted R^2	0.222	0.315	0.055
Residual Std. Error (df = 32)	0.186	0.174	0.205
F Statistic (df = 4; 32)	3.575**	5.137***	1.521
Note:		*p<0.1; **p<0.05; ***p<0.01	

Appendix C: Detailed Learning Analyses

Table C.1: Uncertainty Task, Alignment Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	64,599,442.000	−4.215	22,025.810**	224.549	−317.798
	(74,116,711.000)	(681.453)	(9,764.194)	(1,490.469)	(575.732)
DET	2,010.840	365.476	−43.682***	−18.874	24.308
	(13,320.940)	(542.170)	(13.997)	(19.981)	(22.300)
L	−902.611**	−9.651	−3,645.800**	−11,543.810***	−22,694.880
	(392.720)	(5.814)	(1,795.276)	(3,953.847)	(17,241.140)
ENTR	4,944.413*	−300.305**	5,029.234**	6,791.516***	14,476.480
	(2,474.329)	(133.254)	(2,291.626)	(2,253.497)	(11,203.650)
Constant	−2,592,584.000	847.001***	5,882.223**	22,076.470***	41,932.710
	(2,965,692.000)	(94.213)	(2,510.336)	(7,722.401)	(32,184.850)
Observations	38	40	40	40	40
R^2	0.169	0.804	0.330	0.221	0.092
Adjusted R^2	0.069	0.782	0.253	0.132	−0.011
F Statistic	1.33 (df = 4; 26)	28.80*** (df = 4; 28)	3.44* (df = 4; 28)	1.98 (df = 4; 28)	0.71 (df = 4; 28)

Note: Error df adjusted by 7. *p<0.05; **p<0.01; ***p<0.001

Table C.2: Uncertainty Task, Coordination Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	342,603.100	715.499	−632.505	−637.926**	−53.343
	(517,871.500)	(595.900)	(1,059.251)	(262.715)	(152.182)
DET	11,673.210**	−66.928	−31.371***	−6.207	−27.930
	(4,438.781)	(466.578)	(9.010)	(13.219)	(24.279)
L	592.021	−21.501	−798.242***	−558.341***	−5,834.327***
	(1,157.441)	(18.841)	(125.942)	(123.676)	(881.008)
ENTR	−4,640.338	−353.301*	2,163.754***	382.333	4,122.151***
	(4,166.528)	(192.169)	(443.043)	(358.755)	(1,303.760)
Constant	−17,010.720	850.344***	2,233.743***	2,591.553***	12,374.660***
	(19,690.500)	(92.520)	(475.455)	(431.550)	(1,950.099)
Observations	38	40	40	40	40
R^2	0.318	0.811	0.731	0.674	0.602
Adjusted R^2	0.235	0.789	0.701	0.637	0.556
F Statistic	3.03* (df = 4; 24)	29.98*** (df = 4; 28)	19.05*** (df = 4; 28)	14.47*** (df = 4; 28)	10.57*** (df = 4; 28)

Note: Error df adjusted by 7. *p<0.05; **p<0.01; ***p<0.001

Table C.3: Uncertainty Task, Baseline Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	7,444.118	−31.324	111.169	−161.777	90.318
	(96,136.120)	(801.954)	(1,102.489)	(163.387)	(75.539)
DET	12,744.770***	288.527	−40.226***	−25.647***	−38.917**
	(3,804.908)	(674.001)	(10.952)	(8.612)	(15.340)
L	317.731	−1.841	−66.794	−64.934	−2,533.705***
	(848.454)	(4.866)	(53.208)	(50.545)	(442.410)
ENTR	−4,090.219	−332.130**	535.494	17.006	1,691.631**
	(3,163.603)	(143.626)	(350.384)	(262.940)	(684.850)
Constant	−4,384.431	851.719***	2,334.007***	1,276.885***	5,665.696***
	(3,351.312)	(117.637)	(418.710)	(290.012)	(902.178)
Observations	40	40	40	40	40
R^2	0.324	0.695	0.607	0.507	0.557
Adjusted R^2	0.246	0.660	0.562	0.451	0.506
F Statistic (df = 4; 28)	3.34*	15.96***	10.81***	7.20***	8.79***

Note: Error df adjusted by 7. *p<0.05; **p<0.01; ***p<0.001

Table C.4: Map Task Path Deviation, Alignment Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	−1,284.483*	−175.082*	553.550	37.448	3.631
	(717.451)	(93.601)	(383.045)	(42.326)	(13.058)
DET	−44.763	1,125.770*	−0.231	1.404	−1.771*
	(84.224)	(610.371)	(0.681)	(0.962)	(1.005)
L	0.732	−0.308	132.923*	317.607*	45.496
	(2.981)	(2.379)	(73.975)	(166.371)	(225.427)
ENTR	1.016	33.436	−133.695	−213.428*	2.457
	(20.337)	(35.053)	(100.969)	(109.590)	(150.105)
Constant	79.637	−1,065.386**	−217.702**	−621.720*	−66.438
	(67.343)	(490.389)	(97.466)	(316.809)	(425.252)
Observations	124	124	124	123	124
R^2	0.034	0.104	0.051	0.047	0.033
Adjusted R^2	0.002	0.074	0.019	0.015	0.0003
F Statistic	1.03 (df = 4; 116)	3.37* (df = 4; 116)	1.57 (df = 4; 116)	1.42 (df = 4; 115)	0.98 (df = 4; 116)

Note: Error df adjusted by 3. *p<0.05; **p<0.01; ***p<0.001

Table C.5: Map Task Path Deviation, Coordination Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	−19,994.240	−350.508**	74.398	2.182	−11.138**
	(19,019.810)	(148.062)	(67.499)	(11.618)	(4.672)
DET	13.040	−44.564	0.663	2.058**	0.637
	(111.939)	(290.890)	(0.762)	(0.860)	(1.532)
L	−45.842	8.728	13.527	18.158	429.051***
	(77.712)	(8.074)	(19.707)	(19.066)	(107.782)
ENTR	22.860	73.311	1.012	−33.536	−312.603***
	(141.768)	(70.985)	(68.451)	(37.888)	(110.876)
Constant	884.764	−12.985	−125.791**	−84.114*	−710.774***
	(802.325)	(121.297)	(59.964)	(50.102)	(189.245)
Observations	124	124	124	124	124
R^2	0.030	0.078	0.081	0.099	0.154
Adjusted R^2	−0.002	0.047	0.050	0.069	0.125
F Statistic (df = 4; 116)	0.90	2.44	2.54*	3.19*	5.25***

Note: Error df adjusted by 3.

*p<0.05; **p<0.01; ***p<0.001

Table C.6: Map Task Path Deviation, Baseline Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	−13,364.370	−207.662**	101.056	−3.596	−2.953*
	(8,630.539)	(101.587)	(63.446)	(2.475)	(1.769)
DET	18.498	2,015.435**	0.859	1.838***	0.089
	(111.440)	(895.708)	(0.636)	(0.649)	(1.146)
L	−116.996	−0.595	−0.857	2.832	111.017***
	(72.434)	(1.027)	(0.540)	(3.128)	(34.105)
ENTR	136.929	5.998	−11.692	1.723	−58.301
	(108.985)	(22.270)	(22.953)	(24.544)	(43.988)
Constant	706.306*	−1,826.442**	−66.675	−56.961**	−206.579***
	(369.594)	(778.924)	(46.474)	(24.084)	(68.716)
Observations	124	124	124	124	124
R^2	0.041	0.085	0.067	0.084	0.086
Adjusted R^2	0.009	0.054	0.035	0.054	0.055
F Statistic (df = 4; 116)	1.25	2.68**	2.08	2.67*	2.73*

Note: Error df adjusted by 3. *p<0.05; **p<0.01; ***p<0.001

Table C.7: Map Task Completion Time, Alignment Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	591.522	−175.082*	−236.282	−62.463	−16.023
	(2,174.316)	(93.601)	(1,142.256)	(128.282)	(40.643)
DET	26.419	1,125.770*	3.585*	−6.736**	2.385
	(258.537)	(610.371)	(2.039)	(2.950)	(3.155)
L	−6.591	−0.308	−1,020.097***	−1,845.613***	−2,466.553***
	(9.467)	(2.379)	(223.237)	(509.775)	(694.577)
ENTR	128.162**	33.436	1,317.317***	1,234.506***	1,584.716***
	(61.778)	(35.053)	(303.688)	(333.656)	(462.592)
Constant	−257.278	−1,065.386**	1,093.207***	3,546.359***	4,552.480***
	(204.322)	(490.389)	(296.055)	(971.482)	(1,309.179)
Observations	122	124	122	121	122
R^2	0.117	0.104	0.174	0.126	0.111
Adjusted R^2	0.087	0.074	0.145	0.095	0.081
F Statistic	3.79** (df = 4; 114)	3.37* (df = 4; 116)	5.99*** (df = 4; 114)	4.06** (df = 4; 113)	3.57** (df = 4; 114)

Note: Error df adjusted by 3. *p<0.05; **p<0.01; ***p<0.001

Table C.8: Map Task Completion Time, Coordination Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	366,235.100***	692.678	−834.932***	−137.064***	−37.539***
	(51,631.440)	(472.234)	(158.991)	(24.046)	(11.084)
DET	−444.164	814.175	4.068**	−6.368***	−10.634***
	(305.170)	(953.005)	(1.835)	(1.823)	(3.664)
L	354.831*	−11.752	−292.974***	−325.161***	−2,317.280***
	(208.017)	(25.772)	(46.047)	(39.801)	(261.508)
ENTR	−177.399	−388.635*	712.306***	310.942***	2,088.507***
	(377.703)	(231.616)	(158.429)	(78.208)	(264.267)
Constant	−15,077.780***	−153.827	56.127	1,168.237***	4,773.563***
	(2,180.099)	(398.614)	(141.568)	(108.685)	(461.243)
Observations	122	122	122	122	122
R^2	0.339	0.092	0.510	0.612	0.541
Adjusted R^2	0.316	0.061	0.493	0.599	0.526
F Statistic (df = 4; 117)	14.60***	2.88*	29.63***	44.99***	33.64***

Note: Error df adjusted by 3. *p<0.05; **p<0.01; ***p<0.001

Table C.9: Map Task Completion Time, Baseline Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	89,003.180***	−214.702	−912.257***	−8.712	1.284
	(26,923.630)	(301.081)	(169.013)	(6.678)	(4.648)
DET	−382.098	−4,354.742	−1.150	−7.911***	−8.467***
	(350.462)	(2,687.395)	(1.723)	(1.811)	(3.000)
L	−13.584	4.858	1.038	−27.179***	−499.328***
	(227.412)	(3.107)	(1.491)	(8.415)	(91.468)
ENTR	267.816	−97.559	99.817	2.412	359.960***
	(339.816)	(67.185)	(63.098)	(67.025)	(115.211)
Constant	−3,502.094***	4,614.137*	218.836*	380.759***	1,326.835***
	(1,156.031)	(2,340.077)	(129.415)	(67.633)	(187.865)
Observations	122	122	122	122	122
R^2	0.132	0.220	0.331	0.365	0.392
Adjusted R^2	0.102	0.193	0.309	0.343	0.372
F Statistic (df = 4; 114)	4.33**	8.02***	14.13***	16.36***	18.41***

Note: Error df adjusted by 3. *p<0.05; **p<0.01; ***p<0.001

Table C.10: CSAR Task, Alignment Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	9,093,951.000***	−489.372*	−466.619	−152.081	15.303
	(2,840,001.000)	(249.752)	(1,153.003)	(127.793)	(44.009)
DET	−589.081	6,914.397*	1.668	−1.363	−1.745
	(597.206)	(3,868.915)	(1.098)	(1.591)	(1.947)
L	19.036	−0.748	−76.154	−109.544	−1,257.326*
	(15.104)	(0.633)	(69.296)	(76.914)	(696.793)
ENTR	−56.181	16.769	22.553	98.030*	847.326*
	(109.267)	(62.403)	(79.622)	(50.295)	(428.558)
Constant	−363,231.700***	−6,514.778*	111.844	264.803	2,337.489*
	(113,473.100)	(3,655.477)	(126.862)	(162.525)	(1,341.870)
Observations	116	116	116	114	116
R^2	0.126	0.094	0.037	0.048	0.049
Adjusted R^2	0.095	0.061	0.003	0.013	0.015
F Statistic	2.97* (df = 4; 82)	2.11 (df = 4; 82)	0.79 (df = 4; 82)	1.01 (df = 4; 80)	0.38 (df = 4; 82)

Note: Error df adjusted by 29. *p<0.05; **p<0.01; ***p<0.001

Table C.11: CSAR Task, Coordination Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	93,085.180*	−133.336	−203.272	−50.274*	−11.958
	(49,599.390)	(269.166)	(145.605)	(29.358)	(16.674)
DET	−365.496	87.982	0.081	−0.531	−0.440
	(343.518)	(1,267.024)	(2.011)	(1.111)	(2.789)
L	−34.037	−1.085	−5.328	−0.675	−80.881**
	(30.159)	(1.738)	(7.993)	(1.059)	(37.103)
ENTR	217.029	20.152	−45.642	−3.898	−13.079
	(137.011)	(84.561)	(54.908)	(45.941)	(101.647)
Constant	−3,580.603*	−53.277	151.514	110.293**	327.555**
	(1,908.562)	(1,053.201)	(136.376)	(44.140)	(141.916)
Observations	116	116	116	116	116
R^2	0.055	0.063	0.074	0.069	0.061
Adjusted R^2	0.021	0.030	0.041	0.035	0.027
F Statistic (df = 4; 82)	1.20	1.39	1.64	1.51	1.32

Note: Error df adjusted by 29. *p<0.05; **p<0.01; ***p<0.001

Table C.12: CSAR Task, Baseline Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	34,693.010	−273.854	−21.784	−12.954	−12.333
	(23,134.800)	(186.971)	(58.776)	(11.615)	(11.372)
DET	172.923	1,932.396	−1.557	−1.318	−0.547
	(255.357)	(3,194.002)	(1.331)	(0.883)	(1.691)
L	7.971	−0.448*	−0.243	0.331	−2.491
	(21.948)	(0.263)	(0.553)	(0.869)	(5.952)
ENTR	−75.177	60.182	22.054	13.120	−39.918
	(98.381)	(38.003)	(36.896)	(31.834)	(55.060)
Constant	−1,423.284*	−1,890.598	115.977	84.925**	161.304*
	(853.752)	(3,022.162)	(103.234)	(36.273)	(88.272)
Observations	116	116	116	115	116
R^2	0.033	0.099	0.071	0.086	0.033
Adjusted R^2	−0.002	0.066	0.037	0.053	−0.001
F Statistic	0.69 (df = 4; 82)	2.24 (df = 4; 82)	1.57 (df = 4; 82)	1.90 (df = 4; 81)	0.71 (df = 4; 82)

Note: Error df adjusted by 29. *p<0.05; **p<0.01; ***p<0.001

Table C.13: Diapix Task, Alignment Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	−8,790.082	−336.908	2,314.378**	92.041	−12.844
	(35,883.720)	(221.024)	(866.774)	(124.633)	(21.449)
DET	616.119	2,545.339	−2.356	0.459	3.447
	(968.580)	(3,707.711)	(1.470)	(1.534)	(2.411)
L	0.202	8.167*	−484.904**	−322.465	−984.905
	(25.676)	(4.489)	(189.835)	(291.970)	(829.980)
ENTR	9.155	−146.250	521.386**	131.983	568.189
	(205.513)	(89.084)	(207.920)	(185.990)	(526.253)
Constant	−264.836	−2,011.133	766.629**	613.398	1,812.375
	(1,658.014)	(3,498.856)	(294.064)	(565.599)	(1,562.004)
Observations	48	48	48	48	48
R^2	0.073	0.111	0.250	0.082	0.096
Adjusted R^2	−0.013	0.028	0.180	−0.003	0.012
F Statistic (df = 4; 38)	0.75	1.18	3.16*	0.85	1.01

Note: Error df adjusted by 5. *p<0.05; **p<0.01; ***p<0.001

Table C.14: Diapix Task, Coordination Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	−13,959.790	−222.743	274.014**	17.588	−6.768
	(71,591.090)	(666.976)	(132.688)	(31.932)	(8.307)
DET	59.112	353.899	−1.619	−2.281*	6.892*
	(481.605)	(1,774.167)	(1.141)	(1.319)	(3.493)
L	205.604	7.199	−89.827***	−30.613***	−501.526***
	(337.928)	(24.966)	(21.035)	(10.999)	(119.268)
ENTR	−373.042	−12.141	113.878*	2.734	120.970
	(828.135)	(247.105)	(62.498)	(51.210)	(161.135)
Constant	399.031	−276.878	190.349**	136.742**	957.769***
	(2,883.971)	(1,202.887)	(80.957)	(66.961)	(256.505)
Observations	48	48	48	48	48
R^2	0.082	0.023	0.415	0.313	0.339
Adjusted R^2	−0.003	−0.068	0.361	0.249	0.278
F Statistic (df = 4; 88)	0.85	0.22	6.75***	4.32**	4.87**

Note: Error df adjusted by 5. *p<0.05; **p<0.01; ***p<0.001

Table C.15: Diapix Task, Baseline Models after Controlling for Learning.

Level:	Pitch	Rhythm	Morpheme	Word	Syntax
RR	12,912.030	376.049*	95.261	17.268	0.189
	(25,939.470)	(186.611)	(127.856)	(11.972)	(4.811)
DET	−38.446	−4,103.300	−3.476**	−2.717***	1.248
	(541.239)	(3,734.763)	(1.521)	(0.990)	(2.283)
L	−44.456	−0.520	0.747	−6.808*	−112.160**
	(252.605)	(0.697)	(3.094)	(3.996)	(49.523)
ENTR	193.131	−5.322	24.414	23.410	73.226
	(665.363)	(31.457)	(39.967)	(31.214)	(90.175)
Constant	−648.761	3,848.282	171.836**	74.598**	184.328
	(978.681)	(3,553.767)	(84.534)	(33.914)	(127.447)
Observations	48	48	48	48	48
R^2	0.054	0.107	0.232	0.279	0.129
Adjusted R^2	−0.034	0.024	0.161	0.212	0.048
F Statistic (df = 4; 38)	0.54	1.14	2.87*	3.68*	1.40

Note: Error df adjusted by 5. *p<0.05; **p<0.01; ***p<0.001